

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-125929

(P2001-125929A)

(43) 公開日 平成13年5月11日 (2001.5.11)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード* (参考)
G 0 6 F 17/30	1 7 0	G 0 6 F 17/30	1 7 0 F
C 1 2 M 1/00		C 1 2 M 1/00	A
G 0 1 N 33/48		G 0 1 N 33/48	Z
G 0 6 F 3/00	6 5 1	G 0 6 F 3/00	6 5 1 A
			6 5 1 E

審査請求 未請求 請求項の数18 O L 外国語出願 (全 89 頁) 最終頁に続く

(21) 出願番号 特願2000-227459 (P2000-227459)

(22) 出願日 平成12年7月27日 (2000.7.27)

(31) 優先権主張番号 0 9 / 3 6 2 6 4 9

(32) 優先日 平成11年7月27日 (1999.7.27)

(33) 優先権主張国 米国 (US)

(71) 出願人 500014002

インサイト・ファーマスーティカルズ・インコーポレイテッド

INCYTE PHARMACEUTICALS INC.

アメリカ合衆国カリフォルニア州94304・
パロアルト・ポータードライブ 3174

(72) 発明者 アレックス・ジョージ・コレツァー

アメリカ合衆国 カリフォルニア州94560
ニューアーク、リンコナダ・コート、
8260

(74) 代理人 100096817

弁理士 五十嵐 孝雄 (外3名)

最終頁に続く

(54) 【発明の名称】 生体分子配列データのためのグラフィカルビューア

(57) 【要約】

【課題】 コンピュータベースの生体分子配列情報をグラフィック表示するための、方法、媒体、システムを開示する。

【解決手段】 生体分子配列情報は、一般に、本発明に従って種々様々な形式でグラフィック表示することができる。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの1以上の画面に表示される。

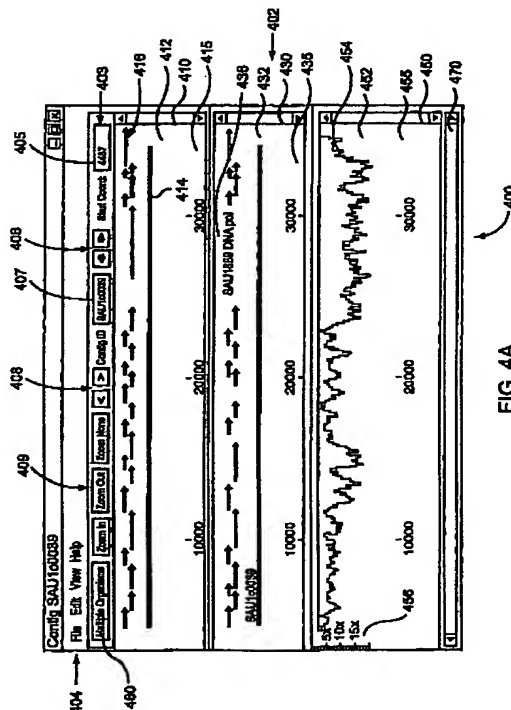


FIG. 4A

BEST AVAILABLE COPY

【特許請求の範囲】

【請求項1】 生体分子配列データを表わすためのコンピュータシステムにおいて実行される方法であって、ユーザからの照会に応じて、データベースから生体分子配列データを検索し、

前記コンピュータシステムのユーザインタフェースにおいて、前記生体分子配列データの要素をグラフィック表示する、ことを備える方法。

【請求項2】 請求項1記載の方法であって、前記グラフィック表示は複数のパネルを備え、前記複数のパネルの少なくとも1つは、前記複数のパネルのうちの少なくとも1つの他のパネルに表示された生体分子配列の一部位についてのカバレッジ情報の深さをグラフィック表示する、方法。

【請求項3】 請求項2記載の方法であって、前記複数のパネルは1つのフレーム内に設けられている、方法。

【請求項4】 請求項3記載の方法であって、前記複数のパネルは、前記生体分子配列データの異なる側面を表わすグラフィック表示を提供する、方法。

【請求項5】 請求項4記載の方法であって、前記生体分子配列データは遺伝子座位データである、方法。

【請求項6】 請求項5記載の方法であって、前記複数のパネルは3つのパネルである、方法。

【請求項7】 請求項6記載の方法であって、前記3つのパネルは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示する第1のパネルと、前記第1のパネルに表示された前記コンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示する第2のパネルと、前記第2のパネルに表示された前記コンティグの一部位のカバレッジ情報の深さをグラフィック表示する第3のパネルと、を備える、方法。

【請求項8】 請求項7記載の方法であって、前記第3のパネルは、前記第2のパネルに表示された前記コンティグの一部位のカバレッジ情報の深さをグラフィック表示する、方法。

【請求項9】 請求項1記載の方法であって、前記方法はJavaプログラミング言語で実行される、方法。

【請求項10】 生体分子配列データを表わすためのコンピュータシステムにおいて実行される方法であって、ユーザからの照会に応じて、データベースから複数のホモログ座位の生体分子配列データを検索し、前記コンピュータシステムのユーザインタフェースにおいて、少なくとも幾つかの前記ホモログ座位をグラフィック表示する、ことを備える方法。

【請求項11】 請求項10記載の方法であって、前記グラフィック表示は1つのパネルを備える、方法。

【請求項12】 コンピュータシステムであって、生体分子配列データを含むデータベースと、前記生体分子配列データに関する照会を受信し、前記照会の結果をグラフィック表示することが可能なユーザインタフェースと、を備えるコンピュータシステム。

【請求項13】 請求項12記載のシステムであって、前記グラフィック表示は複数のパネルを備え、前記複数のパネルの少なくとも1つは、前記複数のパネルのうちの少なくとも1つの他のパネルに表示された生体分子配列の一部位についてのカバレッジ情報の深さをグラフィック表示する、システム。

【請求項14】 請求項13記載のシステムであって、前記複数のパネルは1つのフレーム内に設けられている、システム。

【請求項15】 請求項14記載のシステムであって、前記複数のパネルは、前記生体分子配列データの異なる側面を表わすグラフィック表示を提供する、システム。

【請求項16】 請求項15記載のシステムであって、前記生体分子配列データは遺伝子座位データである、システム。

【請求項17】 請求項16記載のシステムであって、前記遺伝子座位データは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示する第1のパネルと、前記第1のパネルに表示された前記コンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示する第2のパネルと、前記第2のパネルに表示された前記コンティグの一部位のカバレッジ情報の深さをグラフィック表示する第3のパネルと、を備える3つのパネルに表示される、システム。

【請求項18】 生体分子配列データをグラフィック表示するようにアレンジされたプログラム命令を含むコンピュータ読み取り可能な媒体であって、ユーザからの照会に応じて、コンピュータシステムデータベースから生体分子配列データを検索し、前記コンピュータシステムのユーザインタフェースにおいて、前記生体分子配列データの要素をグラフィック表示する、ためのプログラム命令を含むコンピュータ読み取り可能な媒体。

【発明の詳細な説明】

【0001】

【発明の背景】本発明は、概して生命情報学の分野に関し、特に、コンピュータベースの生体分子配列情報をグラフィック表示するための方法、媒体、およびシステムに関する。

【0002】情報科学は、コンピュータおよび統計学的技術を情報管理に応用する学問である。生命情報学は、生体分子配列情報（例えば核酸やタンパク質等）のコンピュータデータベースを素早く検索する方法や、生体分子配列情報を分析して表示する方法、それにDNA配列データからタンパク質の配列、構造、および機能を予測

する方法の開発を含む。

【0003】分子生物学は、実験室主体からコンピュータ主体へと急速に移行しつつある。今日の研究者にとって、配列と表現型の関係の探求には、計量分析やデータベース比較、計算アルゴリズムを高度に駆使できることが必須である。このため、遺伝子の配列や発現、そして分子構造の探求において、コンピュータ資源の使用を避けて通ることは不可能である。

【0004】生命情報学の用途の1つは、生体のゲノムを研究することによって、その遺伝子の配列および配置を決定すること、ならびにその遺伝子が、同種ゲノム内の他の配列や遺伝子または他種の生体内の遺伝子に対して有する関連性を決定することである。こういった情報には、例えば薬効評価や薬物耐性評価へ応用するために、生物医学研究や薬学研究において大きな関心を寄せられている。ゲノム情報操作の実施および理解を容易にするために、高性能なコンピュータデータベースシステムも開発されている。米国カリフォルニア州パロアルトにあるインサイト製薬 (Incyte Pharmaceuticals, Inc.) はそのようなデータベースをいくつか開発しており、その中には、ゲノム配列データが、電子的に記録され、また、公衆の配列データベースから得られる情報で注釈されたものもある。こういった公衆の配列データベースの例として、GenBank (NCBI) や SWISSPROT が挙げられる。結果として得られた情報は、リレーショナルデータベースに格納されて、同種のゲノム内または複数種のゲノム間で、配列と遺伝子の関連性を決定するのに利用される。

【0005】遺伝データの処理や、インサイト製薬が開発したようなリレーショナルデータベースシステムが、遺伝情報の分析に多大な能力と柔軟性を提供している一方で、これらのシステムにさらなる改良を加えることにより、多数の用途において生物学的研究が加速されるであろう。

【0006】対象となる分野の1つに、生体分子配列情報の表示が挙げられる。上述したように、ゲノム研究の重要な目的の1つは、生体遺伝子の配列および配置を決定すること、ならびにその遺伝子が、同種ゲノム内の他の配列や遺伝子、他種の生体内の遺伝子、または関連のタンパク質配列に対して有する関連性を決定することにある。与えられた1種または複数種の生体に対し、その遺伝子座位情報を明確且つ有効に表示することができれば、この課題に大きく貢献することができるに違いない。

【0007】このように、与えられた1種または複数種の生体についての遺伝子座位情報および/または他の生体分子配列情報を、ユーザが明確かつ有効に表示させることのできる表示ツールの開発が望まれている。

【0008】

【発明の概要】本発明は、コンピュータベースの生体分

子配列情報をグラフィック表示するための、方法、媒体、システムを提供することにより、上記のニーズを満たすものである。生体分子配列情報は、一般に、本発明に従って種々様々な形式でグラフィック表示することができる。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの1以上の画面に表示される。

【0009】本発明のグラフィカルビューアは複数のパネルを備え、各パネルは対象とする生体分子配列データに関する情報をそれぞれ異なる方式で1画面または1ページ上に表示することが好ましい。第1のパネルは、例えば、生体分子配列全体またはその生体分子配列のうち対象とする一部位を、対象とする特定の部分配列の位置とともに、グラフィック表示することができる。第2のパネルは、例えば、第1のウィンドウに示された配列全体またはその選択部位を、より詳細にグラフィック表示することができ、これによって、ユーザは、対象とする特定部分配列に焦点を合わせることができる。この第2のパネルは、対象とした特定部分配列に関する注釈等の、付加的な情報を示すことができる。第3のパネルは、例えば、1以上の他パネルに示された生体分子配列データ等の信頼度や発生源を図示した情報を示すことができる。同じ画面または付加的な画面上の付加パネルは、例えば、1以上の他パネルに示された、選択された対象部分配列またはこれに関連する実際のヌクレオチド配列もしくはアミノ酸配列や、その生体分子配列データに関連する他の情報を、表示することができる。

【0010】好ましい一実施形態において、本発明のグラフィカルビューアは、ある生体ゲノムの全体または選択部位を、その個々の座位とともにグラフィック表示することができる。このビューアによって、ユーザは、対象とする特定領域または特定座位に焦点を合わせ、それらを注釈等の付加情報とともにグラフィック表示することができるようになる。また、ビューアに示された配列領域に対し、その配列カバレッジをグラフィック表示できるようにすることもできる。

【0011】このビューアを使用すると、同種の生体ゲノムの他部位から得られる関連座位を表示することもでき、さらに、実際のヌクレオチド配列または詳細な注釈等の、その座位に関する情報を、関連のリレーショナルデータベースシステムから検索することもできる。また、本発明のグラフィカルビューアを使用すると、1つまたはそれ以上の生体ゲノムの複数部位を、対象座位ならびに対応するパラログおよびホモログ (別の生体ゲノムからの関連座位) に基づいて、グラフィック表示および比較することもできる。

【0012】本発明の好ましい一実施形態に従ったグラフィカルビューアは、対象とするゲノムデータを、その

5

ゲノムデータに関して各々が異なる情報を表示するような複数のパネルを使用して、グラフィック表示することが好ましい。本発明の特に好ましい実施形態において、グラフィカルビューアは、1つの画面上にメインパネルを3つ有する。すなわち、考慮中のゲノム全体を示すレジェンドビューアと、特に関心のあるゲノム部位の領域にユーザが焦点を合わせる（ズームインする）ことを可能とするターゲットビューアと、考慮中のゲノム部位の長さに対してカバレッジの深さを示したグラフィック情報を含む配列深さビューアと、の3つである。

【0013】一態様において、本発明は、生体分子配列データを示すためのコンピュータシステムで実行される方法を提供する。この方法は、ユーザの照会に応じてデータベースから生体分子配列データを検索し、その生体分子配列データのエレメントをコンピュータシステムのユーザインタフェース内にグラフィック表示することを含む。この際、生体分子配列データを様々な側面から表示する複数のパネルが単一フレーム内に含まれているようにしてもよい。

【0014】好ましい一実施形態において、生体分子配列データは、遺伝子座位データを含み、3つのパネル上にグラフィック表示される。すなわち、第1のパネルは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示し、第2のパネルは、第1のパネル上に示されたコンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示し、第3のパネルは、第2のパネル上に示された配列データを決定するために実施されたシークエンス操作の回数を示す情報をグラフィック表示する。第3のパネルは、第2のパネル上に示されたコンティグの一部位を構成するのに使用される配列か、または第2のパネル上に示されたコンティグの一部位に対するカバレッジ情報の深さを、グラフィック表示することが可能である。

【0015】別の態様において、本発明は、生体分子配列データを示すためのコンピュータシステムで実行される別の方法を提供する。この方法は、ユーザの照会に応じて、データベースから複数のホモログス座位の生体分子配列データを検索し、コンピュータシステムのユーザインタフェースにおいて、少なくとも幾つかのホモログス座位をグラフィック表示することを含む。

【0016】さらに別の態様において、本発明はコンピュータシステムを提供する。このコンピュータシステムは、生体分子配列データを含むデータベースと、ユーザインタフェースと、を備える。このユーザインタフェースは、生体分子配列データに関連する照会を受信し、その照会結果をグラフィック表示することができる。

【0017】さらにまた別の態様において、本発明は、生体分子配列データをグラフィック表示するようにアレンジされたプログラム命令を含むコンピュータ読み取り可能な媒体を提供する。このコンピュータ読み取り

6

可能な媒体は、ユーザの照会に応じて、データベースから生体分子配列データを検索し、コンピュータシステムのユーザインタフェースにおいて、その生体分子配列データのエレメントをグラフィック表示するためのプログラム命令を含む。

【0018】以下の説明と、本発明の基本原理を実施例に従って示す明細書と添付の図面とから、本発明の上述したおよびその他の特徴および利点が一層明らかとなる。

10 【0019】

【好ましい実施形態の詳細な説明】次に、本発明の好ましい実施形態について詳しく言及する。添付の図面には、好ましい実施形態が例示されている。以下では、本発明を好ましい実施形態と関連付けて説明するが、これは、本発明を1つまたはそれ以上の好ましい実施形態に限定することを意図するものではない。逆に、添付の特許請求の範囲によって定義される本発明の精神および趣旨の範囲内に含まれ得る、代替物、変更態様、および均等物をもカバーするものである。以下の説明では、本発明の完全な理解を促すために多くの項目を特定しているが、本発明は、これらの項目の一部または全てを特定しなくても実施することが可能である。そのほか、本発明を不必要に不明瞭化するのを避けるため、周知の処理工程の説明は省略した。

【0020】導入：本発明は、コンピュータベースの生体分子配列情報をグラフィック表示するための、方法、媒体、システムを提供する。一般に生体分子配列情報は、本発明に従って種々様々な形式でグラフィック表示することが可能である。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの1以上の画面に表示される。

【0021】本発明のグラフィカルビューアは複数のパネルを備え、各パネルは対象とする生体分子配列データに関する情報をそれぞれ異なる方式で1画面または1ページ上に表示することが好ましい。第1のパネルは、例えば、生体分子配列全体またはその生体分子配列のうち対象とする一部位を、対象とする特定の部分配列の位置とともに、グラフィック表示することができる。第2のパネルは、例えば、第1のウィンドウに示された配列全体またはその選択部位を、より詳細にグラフィック表示することができ、これによって、ユーザは、対象とする特定部分配列に焦点を合わせることができる。この第2のパネルは、対象とした特定部分配列に関する注釈等の、付加的な情報を示すことができる。第3のパネルは、例えば、1以上の他パネルに示された生体分子配列データ等の信頼度や発生源を図示した情報を示すことができる。同じ画面または付加的な画面上の付加パネル

7

は、例えば、1以上の他パネルに示された、選択された対象部分配列またはこれに関連する実際のヌクレオチド配列もしくはアミノ酸配列や、その生体分子配列データに関連する他の情報を、表示することができる。

【0022】好ましい一実施形態において、本発明のグラフィカルビューアは、ある生体ゲノムの全体または選択部位を、その個々の座位とともにグラフィック表示することができる。このビューアによって、ユーザは、対象とする特定領域または特定座位に焦点を合わせ、それらを注釈等の付加情報とともにグラフィック表示することができるようになる。また、ビューアに示された配列領域に対し、その配列カバレッジをグラフィック表示できるようにすることもできる。

【0023】このビューアを使用すると、同種の生体ゲノムの他部位から得られる関連座位を表示することもでき、さらに、実際のヌクレオチド配列または詳細な注釈等の、その座位に関する情報を、関連のリレーショナルデータベースシステムから検索することもできる。また、本発明のグラフィカルビューアを使用すると、1つまたはそれ以上の生体ゲノムの複数部位を、対象座位ならびに対応するパラログおよびホモログ（別の生体ゲノムからの関連座位）に基づいて、グラフィック表示および比較することもできる。

【0024】本発明の好ましい一実施形態に従ったグラフィカルビューアは、対象とするゲノムデータを、そのゲノムデータに関して各々が異なる情報を表示するような複数のパネルを使用して、グラフィック表示することが好ましい。本発明の特に好ましい実施形態において、グラフィカルビューアは、1つの画面上にメインパネルを3つ有する。すなわち、考慮中のゲノム全体を常に示すレジェンドビューアと、特に関心のあるゲノム部位の領域にユーザが焦点を合わせる（ズームインする）ことを可能とするターゲットビューアと、考慮中のゲノム部位の長さに対してカバレッジの深さを示したグラフィック情報を含む配列深さビューアと、の3つである。

【0025】上述したように、本発明のグラフィカルビューアは、本発明の好ましい実施形態に関連して説明する遺伝子座位情報以外の生体分子配列情報の表示にも、もちろん使用することが可能である。本発明のグラフィカルビューアは、例えばペプチドまたはヌクレオチドの配列情報を表示するのに使用しても良く、例えばBLASTまたはFASTAを検索して配列を比較した結果として得られる、実際の配列を表示するのに使用することも可能である。

【0026】グラフィカルビューアの環境：上述したように、本発明のグラフィカルビューアは、例えば、米国カリフォルニア州パロアルトにあるインサイト製薬（Incyte Pharmaceuticals, Inc.）によって開発され、また、米国特許第5,970,500号、第5,953,727号、第5,966,712号、第6,023,6

8

59号などに記載されている生体分子配列のリレーショナルデータベースシステムとともに使用することが好ましい。本発明のグラフィカルビューアによって表示されるデータは、当業者が周知の技法およびコマンドを使用することにより、このようなデータベースから得られるものである。図1Aおよび図1B、そして以下の関連する説明から、本発明のグラフィカルビューアの動作状況が分かる。

【0027】図1Aは、リレーショナルデータベースにおいて情報を格納および検索するのに適したネットワークシステム130を示す図であり、本発明のグラフィカルビューアをサポートするのに適したシステムである。ネットワーク130は、ネットワークサーバ136およびクライアント138a、138b（より多数のクライアントの代表である）と、これらを接続するネットワークケーブル134とを備えている。このケーブル134は、ファイアウォールゲートウェイ140にも接続されており、ファイアウォールゲートウェイ140は、さらにインターネット142に接続されている。

【0028】ネットワーク130は、ローカルエリアネットワーク（LAN）や広域ネットワーク（WAN）を含む周知の従来型ネットワークシステム（例えば、イーサネット（登録商標）やIBMトークンリング等を使用したもの）のうちのいずれであっても良い。このネットワークは、周知のフォーマット（例えばURL）のクライアントコールを任意のパラメータ情報とともに、ケーブルまたはワイヤ134を介して伝送するのに適した（1以上のパケットの）フォーマットにパッケージ化し、データベースサーバ136へ送達する機能を備えている。

【0029】サーバ136は、ソフトウェアを実行させるのに必要なハードウェアを備えている。このソフトウェアは、（1）ユーザリクエストを処理するためのデータベースのデータにアクセスし、（2）クライアントマシン138a、138bに情報を与えるためのインタフェースを提供する。図1Aの好ましい一実施形態において、サーバマシン上で実行されるソフトウェアは、サーバクライアント間でページデータを提供するためのワールドワイドウェブ（WWW）プロトコルをサポートしている。本実施形態において、URLおよびHTTP機能を有するウェブサーバ156は、HTTPプロトコルを介してクライアントとの通信を行う。

【0030】クライアント/サーバ環境、データベースサーバ、リレーショナルデータベース、およびネットワークは、技術文献や、営業用文献、特許文献に詳細に記載されている。リレーショナルデータベースおよびクライアント/サーバ環境に関しては広く議論されており、特に、SQLサーバに関するデータベースサーバに関しては、例えば、「Nath, A., The Guide To SQL Server, 2nd ed., Addison-Wesley Publishing Co., 1995」を

参照すると良い。

【0031】図示するように、サーバ136は、リレーショナルデータベース管理システム152と、ワールドワイドウェブアプリケーション154と、ワールドワイドウェブサーバ156とを実行させるオペレーティングシステム150（例えばUNIX（登録商標））を備えている。サーバ136上のソフトウェアは種々の構成を想定でき、例えば、1つのマシン上に準備されていてもよいし、複数のマシンに分散して準備されていてもよい。

【0032】ワールドワイドウェブアプリケーション154は、データベース言語ステートメント（例えば、スタンダードクエリーラングージ（SQL）ステートメント）の生成に必要な実行可能コードを含む。一般に、実行可能コードは埋め込みSQLステートメントを含む。また、アプリケーション154は、コンフィギュレーションファイル160を含んでいる。このコンフィギュレーションファイル160は、サービスユーザリクエストにアクセスされていなければならない種々の外部および内部データベースと、サーバを構成する種々のソフトウェアエンティティと、に対するポインタおよびアドレスを含んでいる。また、コンフィギュレーションファイル160は、サーバ資源へのリクエストを適切なハードウェアに直接導く。これは、サーバが2以上の個別のコンピュータに分散している場合に必要となる。

【0033】各クライアント138a、138bは、サーバ136へのユーザインタフェースを提供するためのワールドワイドウェブブラウザと、HTMLページを生成するのに必要なコードとを備えている。各クライアント138a、138bは、ウェブブラウザを介して、例えば配列データベース144および/またはゲノムデータベース146からデータ検索するための検索リクエストを作成する。ユーザは、通常、ボタン、プルダウンメニュー、スクロールバー等の従来からグラフィカルユーザインタフェースとして採用されているユーザインタフェースエレメントを、指示したりクリックしたりする。クライアントのウェブブラウザでこのように作成されたリクエストは、ウェブアプリケーション154に伝送される。ウェブアプリケーション154は、リクエストのフォーマットを設定し、配列データベース144またはゲノムデータベース146から適切な情報を抽出するのに利用される照会を生成する。

【0034】図中に示されている実施形態において、ウェブアプリケーションは、先ずデータベース言語（例えばSybaseまたはOracle SQL）で照会を作成することによって、ゲノムデータベース146のデータにアクセスする。このとき、このデータベース言語の照会が、リレーショナルデータベース管理システム152に引き渡され、そこで、データベース146から関連の情報を抽出できるように処理される。リクエストが

配列データベース144にアクセスするものである場合、ウェブアプリケーション154は、データベース管理システム152のサービスを利用することなく直接そのデータベースにリクエストを通信する。

【0035】ユーザリクエストにサービスする手順を、図1Bにさらに示す。本実施形態において、ワールドワイドウェブサーバ、および/または、サーバ136の実行可能なウェブアプリケーションコンポーネントは、クライアントマシンにハイパーテキストマークアップ言語文書（HTMLページ）164を提供する。HTML文書は、クライアントマシンにおいて、ユーザがデータベース146にアクセスするためのリクエストを作成するのに利用されるユーザインタフェース166を提供する。このリクエストは、サーバ136のウェブアプリケーションコンポーネントによってSQL照会168に変換される。この照会は、サーバ136のデータベース管理システムコンポーネントによって使用されて、データベース146内の関連のデータにアクセスし、そのデータを適切なフォーマットでサーバ136に提供する。そして、サーバ136は、ウェブアプリケーション154を介して新しいHTML文書を生成し、データベース情報を、ユーザインタフェース166におけるビューアとしてクライアントに転送する。

【0036】図1Aで示される実施形態では、サーバ136とクライアント138a、138bの通信にワールドワイドウェブサーバおよびワールドワイドウェブブラウザを利用しているが、他の通信プロトコルを用いてもよい。例えばクライアントコールは、SQL変換するウェブアプリケーション154に依らずに、SQLステートメントとして直接パッケージ化されていてもよい。クライアントは、クライアントブラウザを使用せず、直接データベースに照会することも可能である。

【0037】ネットワーク130がワールドワイドウェブサーバおよびクライアントを利用する場合には、TCP/IPプロトコルをサポートしていなければならない。このようなローカルネットワークは、イントラネットと呼ばれることもある。このようなイントラネットの利点は、ワールドワイドウェブ上に属するパブリックドメインデータベース（例えばGenBankワールドワイドウェブサイト）と容易に通信できることにある。本発明の特に好ましい実施形態において、クライアント138a、138bは、ウェブブラウザおよびウェブサーバ156によって提供されたHTMLインタフェースを使用して、イントラネットデータベース上に属するデータに（例えばハイパーテキストリンクを介して）直接アクセスすることができる。

【0038】ローカルデータベースの内容をプライベートに維持したい場合は、ファイアウォール140は配列データベース144およびゲノムデータベース146の内容を秘密に保たなくてはならない点に留意する必要がある

ある。

【0039】また、配列データベース144は、異なる種から得られるゲノム配列の単一ファイルを有したフラットファイルデータベースであるようにしてもよい。他には、配列データを、種別に、あるいは、ローカルデータベースに特有な配列（すなわち、GenBank等の外部データベースで全く見つからなかった配列）であるか否かに基づいて、区分するようにしてもよい。

【0040】ゲノムデータベース146内の情報は、リレーショナルフォーマットで格納されていることが好ましい。このようなリレーショナルデータベースは、関係代数によって定義される1組の操作をサポートする。リレーショナルデータベースは、一般に、同データベース内に含まれる各データに対し、列と行とで構成された複数のテーブルを含む。各テーブルは、テーブル内の行を一意的に識別する値を有した任意の列または列の組である基本キーを有する。リレーショナルデータベースのテーブルはまた、別のテーブルの基本キーの値に一致する値を有した列または列の組である外部キーを有する。リレーショナルデータベースは、一般に、データベース内の関係を管理する関係代数の基礎をなす1組の操作（選択する、射影する、作成する、結合させる、および分割する）に従う。上述したように、リレーショナルデータベースは、周知であり、また、多くの文献がある（例えば、前掲の「Nath, A., The Guide To SQL Server」を参照のこと）。

【0041】リレーショナルデータベースは、種々の方法で提供される。例えばOracle（商標）データベースでは、各テーブルにそれぞれ所有権が明記されたワークスペースがあるので、テーブルが物理的に分離されることはない。対照的に、Sybase（商標）データベースでは、異なるデータベース間でテーブルを物理的に分離させてもよい。

【0042】複数のユーザに対応するネットワーク130の特定の構成では、1つのマシンにゲノムデータベースおよび配列データベースの双方を準備する。大きいな配列を検索する場合は、同サイズの第2のプロセッサを用意して、アプリケーションを2つのマシンに分割し、レスポンス時間を改善することが望ましい。

【0043】デュアルプロセッササーバマシンとしては、Sun-Ultra-Sparc2（商標）（米国カリフォルニア州マウンテンビューにあるサンマイクロシステムズ）や、SGI-Challenge L（商標）（米国カリフォルニア州マウンテンビューにあるシリコングラフィックス）、DEC-2100A（商標）（米国マサチューセッツ州メイナードにあるデジタルエレクトロニクス）などのワークステーションが適している。マルチプロセッサシステム（4プロセッサ以上）としては、Sun-Ultra-Sparc-Enterprise 4000（商標）や、SGI-Challe

nge XL（商標）、DEC-8400（商標）などが適している。サーバマシンは、ネットワーク130のために構成されてTCP/IPプロトコルをサポートしていることが好ましい。

【0044】オペレーティングシステムは、利用されるワークステーションに依存しており、例えばSun-Solaris 5.5（Solaris 2.5）、SGI-IRIX 5.3（またはそれ以上）やDEC-Digital UNIX 3.2D（またはそれ以上）を用いることができる。

【0045】本発明に関連して使用されるデータベースは、4 X 4 Gb+FWSCSI-2、Fiber Link Raid Units 20Gb+、または4 DAT Tape Driveを介してダウンロードすることができる。また、CD-ROMドライブを用いることもできる。

【0046】クライアントマシンとしては、例えば、Macintosh（商標）（米国カリフォルニア州キューバティノーにあるアップルコンピュータ）や、PC、またはUNIXワークステーションを用いることができる。また、クライアントマシンは、Netscape（登録商標）やInternet Explorer Web Browerを伴うTCP/IP対応であるべきである。

【0047】ネットワークは、10Base-Tや、100Base-T、より高速の接続が可能なTCP/IP対応でもよく、外部のデータベースへのHTMLハイパーリンクのためにインターネットへのアクセスを提供する。

【0048】図1Cは、本発明の好ましい一実施形態におけるグラフィカルビューア機能のアクセス可能性を示す図である。本発明のグラフィカルビューアは、ユーザが、生体分子リレーショナルデータベースのユーザインタフェースに表示されたユーザインタフェース画面の集合（例えばHTMLまたはJava（登録商標）ページ）を介して利用できる1組の機能とともに提供されることが好ましい。通常、このインタフェースは、種々の照会ラインが続くメインのビューアページを有している。好ましい一実施形態において、メインビューアページ（およびその他のグラフィカルビューア）は、ネットワークシステム上で実行されるJava（登録商標）ベースのアプレットである。当業者は、ここで説明される機能を与えられれば、本発明によるグラフィカルビューアを、Java（登録商標）またはその他のプログラミング環境に実装することができる。ビューアページは、通常、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベースのユーザインタフェースの一部として提供される、別のページからアクセスされる。

【0049】例えば、ユーザインタフェース画面（例え

ばHTMLページ) 170は、複数の生体分子配列に関するテキスト情報を表示する。ページ170に表示された1以上の配列を、例えば、GUIとして設けられたポインタを使用して選択することにより、その選択された配列に関する付加的な情報を表示する別のページ180にアクセスする。このページ180に設けられたボタンを選択することにより、メインのグラフィカルビューアページ190にアクセスできる。グラフィカルビューアページ(例えばJava(登録商標)ページ)190は、選択された配列に関する情報を図示する。このページには、ユーザによるグラフィカル表示の変更を可能とするための複数のボタン192を設けることが好ましい。ボタン192はさらに、付加的なグラフィカルビューアページ194、196にアクセスするための複数のボタンを備えていても良い。これらのページ194、196は、ページ190にグラフィック表示された配列情報に関する付加的な情報を、グラフィックまたは他の形式で表示する。

【0050】遺伝子座位の提供: 以下では、本発明を、遺伝子座位情報をグラフィック表示するための特に好ましい一実施形態と関連させて説明する。本発明は、米国特許第5,970,500号に記述されるような、菌類データに適したデータベースに関連付けて説明されるが、本発明の適用範囲はこれに限定されない。例えば、本発明は、動物(例えばヒト、霊長類、げっ歯類、両生類、昆虫など)や植物の配列等の他の生体配列データに適したデータベースに関連して使用されるグラフィカルビューアをも範囲に含んでいる。

【0051】上述したように、本発明のグラフィカルビューアは、ユーザが、生体分子リレーショナルデータベースのユーザインタフェースに表示されたユーザインタフェース画面の集合を介して利用できる1組の機能とともに提供されることが好ましい。メインビューアページは、通常、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベース(この場合は菌類ゲノムデータベース)のユーザインタフェースの一部として提供される、別のページからアクセスされる。図2は、菌類ゲノムデータベースから得られたこのような別のページの一例を示す。コンティグ結果ページ200は、菌類の生体(この場合では黄色ブドウ球菌)のゲノム配列のうちの特定の「コンティグ」(重複配列の一集合体)、すなわちコンティグSAU1c0039に局在化した遺伝子の座位(座位IDによって識別される)のリストを表示している。

【0052】ユーザは、コンティグ結果ページ200に含まれる特定の座位IDをクリックすることにより、図3に示されるような、座位情報ページにアクセスすることができる。例えばページ200の座位ID: SAU100241をクリックすると、その座位: SAU100241に関する詳細を表示する座位情報ページ300が返ってくる。この

ページは、選択されたときに本発明のグラフィカルビューアを起動するグラフィカルビューアボタン302を含んでいる。

【0053】図4Aは、座位情報ページ300内のグラフィカルビューアボタン302を選択することによってアクセスされるメインのグラフィカルビューアページ400を示している。この好ましい実施形態において、グラフィカルビューアは、あるコンティグの一部位およびその関連座位のグラフィック表示を提供するJava(登録商標)ベースのアプレットである。本発明のグラフィカルビューアは、複数の個々のコンポーネントビューアを備えていることが好ましい。2以上のコンポーネントビューアが装備されている場合は、それらを1つのフレーム内に表示することにより、グラフィック表示されるデータをユーザに効率的に伝えることができる。好ましい実施形態では、3つのコンポーネントビューアが1つのフレーム内に表示される。

【0054】ページ400のグラフィカルビューア402は、1つの画面に3つのビューアコンポーネントパネルを有する。トップパネル410は、考慮中のゲノム全体を表示するレジェンドビューア412を備える。ミドルパネル430は、ユーザが特に関心のあるゲノム部位の領域に焦点を合わせる(ズームイン)ことを可能とするターゲットビューア432を備える。ボトムパネル450は、ターゲットビューア422に表示されているゲノム部位の長さに対して、カバレッジの深さを示したグラフィック情報を表示する配列深さビューア452を備える。

【0055】グラフィカルビューアページ400は、付加情報へのアクセスおよびその表示のためのボタンおよびウィンドウを、ページ400のトップ403に沿って幾つか有している。また、種々のコマンドおよび制御機能を一覧表示するプルダウンメニューにアクセスするためのメニューバー404が備えられている。各ビューアパネル410のボトムには、スケール415、435、455がそれぞれ示されている。これらの機能の使用方法に関しては、さらに後述する。

【0056】レジェンドビューア412は、ユーザが前の画面で1つのコンティグを選択した際にビューアによってロードされたそのコンティグの全部位を常に示している。好ましい一実施形態において、ビューアは、そのコンティグ配列に関して所定のデフォルト数分の塩基対をロードする。コンティグがデフォルトより短い場合は、コンティグ全体が表示されてデフォルトが調整される。例えば本実施形態において、ビューアは、コンティグ結果画面200上のリストに示される第1の座位、すなわちg2462967(ヒットIDによって識別される)からスタートし、合計30,000の塩基対をロードする。示される塩基対の番号およびコンティグ上のその位置は、レジェンドビューアパネル410のボトムに示され

るスケール 415 をもとに決定される。デフォルト値は、もちろん任意の望ましい数に変更することが可能である。

【0057】レジェンドビューア 412 は、スケール 415 からわかるように、コンティグ SAU1c0039 を、座標（塩基対番号）4467 を起点とし座標 34, 467 に至るライン 414 としてグラフィック表示している。ビューアに示されたコンティグは、コンティグ ID ウィンドウ 407 内で識別される。また、レジェンドビューア 412 によって示されるコンティグ部位の開始座標（すなわち、選択された座位 g2462967 の開始座標：4467）は、開始座標ウィンドウ 405 に示されている。これらのウィンドウ 405, 407 は、以下で説明するように、ビューアによって表示される情報を制御するための情報を入力するのに使用してもよい。ユーザは、方向ボタン 406 をクリックすることにより、レジェンドビューア 412 および他のコンポーネントビューアで、コンティグの上流部位または下流部位を見ることができ

る。

【0058】レジェンドビューア 412 は、コンティグに加え、コンティグの部位上にある種々の座位をも表示する。これらの座位を表示する方法から、本発明のグラフィカルビューアが、効率的な情報表示において力を発揮することがわかる。

【0059】座位は、矢印 416 で表わされる。各座位は、そのコンティグ上の位置およびそれが読まれる方向に従って、コンティグライン 414 に沿って配置されている。矢印は、その座位が読まれる方向を表わしている。前進方向（+）に読まれる座位はコンティグライン 414 の上方に示され、後退方向（-）に読まれる座位はコンティグライン 414 の下方に示される。また、グラフィック表示された座位に関する情報を伝えるために、他のグラフィック機能を使用してもよい。例えば、確定された信頼性の閾値を上回る配列の元となる座位は、破線の矢印で示すことが可能である。

【0060】この好ましい実施形態では、タンパク質機能に基づいて座位を色分けする。タンパク質を種々の機能カテゴリにグループ分けし、各カテゴリに色を割り当てる。例えば、この好ましい実施形態では、座位に対応するタンパク質は以下のカテゴリ／色に従ってグループ分けされる。すなわち、運動性／淡青色、病原性／赤色、輸送性／黄緑色、制御性／マゼンタ、高分子の新陳代謝／黄色、低分子の新陳代謝／濃青色、構造／濃緑色、未分類／黒である。もちろん、他のカテゴリおよび色を使用してもよい。座位の機能を表示するこれら矢印や色分けは、レジェンドビューアと以下で説明するターゲットビューアとの両方で使用される。

【0061】ターゲットビューア 432 は、初めはレジェンドビューア 412 と同じ範囲を表示するが、ズームボタン 409 をクリックすることにより、範囲を変化さ

せることが可能である。「Zoom In」ボタンを使用すると、レジェンドビューア 412 に示されたコンティグの部位を詳細表示することができる。そしてターゲットビューア 432 には、ズーム倍率に応じて調整されたスケール 435 を伴う詳細表示が示される。「Zoom Out」ボタンを使用すると、最大でレジェンドビューアのために選択されたデフォルトの塩基対数（最小倍率）まで、コンティグを広範囲に表示することができる。「Zoom None」ボタンを使用すると、自動的に最小倍率に戻る。

【0062】本発明のグラフィカルビューアによって提供されるレジェンドビューア 412 に示されたコンティグ 414 の対象部位に焦点を合わせるための別の方法としては、コンティグ 414 の対象部位の周囲にカラー（例えば赤色）ボックス等のアウトラインを提供し、それをターゲットビューア 432 に表示する方法が挙げられる。この好ましい実施形態において、赤色ボックスがレジェンドビューアパネル全体を囲んだ場合には、ターゲットビューアは 30, 000 の塩基対を全て表示する。これが、図 4A に示された状態である。上述したように、赤色ボックスはズームボタン 409 を用いて調整される。

【0063】ユーザは、赤色ボックス（ラバーバンディングとして知られる）を直接調整することによっても、コンティグ上の一領域を拡大・縮小することが可能である。赤色ボックスの範囲の変更は、任意のビューアパネルで 1 つの場所をクリックし、カーソルをマウスとともに別の場所にドラッグすることによって行われる。このとき、赤色ボックスがこれら 2 点に挟まれる領域を取り囲み、ターゲットビューアおよび配列深さビューアでは、この領域のみが見られるようになる。図 4B は、ユーザが、レジェンドビューア 412 に示されるコンティグ 414 のうち、座標が約 14, 200 から 18, 200 までの部位 434 を拡大したときのビューア 402 の更新ページを示したものである。ターゲットビューア 432 のボトムにあるスケール 435 は、拡大されたターゲット表示の新しい範囲に応じて調整されている。

【0064】ターゲットビューアの他の機能は、座位が注釈されていることである。図 4A および 4B で見られるように、注釈情報を載せるのに十分な長さの座位の場合は、座位の矢印に注釈 436 が付される。対象座位が注釈を載せるのに短すぎる場合には、ユーザは、注釈のグラフィック表示に十分な長さまで、その座位を拡大すれば良い。

【0065】ターゲットビューア 432 にグラフィック表示された座位をクリックして選択することにより、個々の座位に対して更なる分析をすることができる。選択された座位は、例えばカラー（例えば赤色）ボックスで囲む等の何らかの方法によって強調される。そして、座位の表示をダブルクリックすることによって、その座位

の詳細を見ることができる。つまり、ダブルクリックによって、図 5 A に示されるような、選択オブジェクト詳細ウィンドウが開く。選択オブジェクト詳細ウィンドウ 500 は、座位に関する情報を含んでいる。この情報には、座位 ID、遺伝子（機能別）カテゴリ、塩基対の範囲、例えば GenPept データベース等の他の配列データベースに対する配列のホモログスマッチ（返されるホモログスマッチ数は、例えば上位 5 つ等の予め設定された数に制限されることが好ましい）、検索者にとって有用でかつグラフィカルビューとともに使用されるデータベースシステムの他の機能に関連するその他の情報などが含まれる。ウィンドウ 500 で提供される情報フィールドの多くは、他の HTML ページまたは他の画面へのハイパーリンクであってもよい。

【0066】選択オブジェクト詳細ウィンドウ 500 は、アライメントボタン 502 を備える。このボタンをクリックすると、座位の配列およびそのホモログスマッチをグラフィック表示するアライメントビューにアクセスすることができる。図 5 B は、本発明の好ましい一実施形態におけるアライメントビュー 510 の一例を示したものである。アライメントビュー 510 は、パネルを 3 つ有する。上 2 つのパネル 512、514 は、図 5 A に示される座位（SAU101156）をグラフィック表示する。3 番目のパネル 516 は、図 5 A で記された 5 つのホモログをグラフィック表示する。アライメントビューページは、グラフィック表示の制御に使用されるいくつかのボタン 518 を含んでいる。このページは、特に、下 2 つのパネル 514、516 に示される座位を配列レベルで拡大・縮小するのに使用されるズームボタン 520 を備える（一方、一番上のパネル 512 は、座位全体を表示したままである）。図 5 C は、このようなズーム機能を示したものであり、上のパネル 512 では、下 2 つのパネル 514、516 に、ホモログとともに配列レベルで示された座位の一部が、カラーボックス 522 で囲まれている。この実施形態ではアミノ酸配列が示されているが、他の実施形態では、対応するヌクレオチド配列を示すことも可能である。

【0067】グラフィカルビューページ 400 が有する他の機能のうち、同ページのボトムに設けられたスクロールバー 470 は、ターゲットビュー 432 の表示範囲を、レジェンドビュー 412 に示されるコンティグ配列の一部に焦点合わせした場合に有用となる。ユーザは、スクロールバー 470 をコンティグ 414 の一部 434 に沿って移動させることにより、ターゲットビュー 432 の表示範囲に、同コンティグ 414 の上流または下流の部位を表示させることができる。

【0068】本発明のこの実施形態において、グラフィカルビュー 402 の 3 番目のパネル 450 は、配列深さビュー 452 である。配列深さビュー 452 は、カバレッジの深さ、すなわちコンティグの所定部位がシ

ークエンシングされた回数を、同コンティグの長さに対して示したグラフを表示する。配列深さビュー 452 は、ターゲットビュー 432 に表示されたコンティグまたはコンティグの一部を対象としたグラフを表示する。このため、ターゲットビュー 432 とレジェンドビュー 412 が同じ範囲を示す図 4 A の場合には、配列深さビュー 452 が表示するグラフは、配列深さビューパネル 450 のボトムに位置するスケール 455 が示すように、カバレッジの深さを、コンティグ 414 のうち座標 4467~34、467 に相当する 30、000 塩基対に対して示したものである。一方、図 4 B の場合には、配列深さビュー 452 が表示するグラフは、調整後のスケール 455 が示すように、カバレッジの深さを、コンティグのうちターゲットビューでクローズアップされている座標約 14、200~18、200 に相当する部位 434 の約 4000 塩基対に対して示したものである。配列深さビューは、さらに、グラフに表示されたシークエンシングの通過回数を示す第 2 のスケール 456 を y 軸上に備えている。

【0069】このカバレッジ情報の深さを示す方法から、本発明のグラフィカルビューが効率的に情報表示できることがわかる。グラフィカルビューのユーザは、一見するだけで、ビュー中の他のパネルに示された配列情報の信頼性に関する有用な情報を、素早く知ることができる。本発明のこの好ましい実施形態における配列深さビュー 452 は、カバレッジを配列分布グラフ 454 として表示する。カバレッジ情報の深さをこのように表示すれば、この情報を、y 軸スケール 456 を参照することにより、視覚的に明確な印象を与えかつ容易に定量が可能なグラフ形式で、特に効率的に提供できる。種々の領域のカバレッジデータも、この形式によって容易に比較することができる。

【0070】本発明の別の実施形態では、配列深さビューは、カバレッジ情報の深さを別の方法でグラフィック表示してもよい。例えば、コンティグが構成される実際の配列を表示してもよい。この方法で配列のカバレッジ情報を表示すれば、データ収集プロセスに関心のあるユーザにとって有用な情報、例えばコンティグの形成に使用される情報を提供することができる。

【0071】上述したように、グラフィカルビューページ 400 は、同ページ 400 のトップ 403 に沿って、付加的な情報へのアクセスおよびその表示を目的としたいくつかのボタンを備える。これらのうち、開始座標 405 ウィンドウやコンティグ ID 407 ウィンドウ等の幾つかに関しては、すでに説明済みである。図 6 および 7 は、本発明のグラフィカルビューのこの実施形態における付加的な機能を示したものである。

【0072】開始座標ウィンドウ 605 は、レジェンドビュー 612 に表示されたコンティグ配列の開始座標を表示するのに加え、ユーザから異なる開始座標の入力

を受け付けることも可能である。異なる開始座標を入力すると、コンティグ配列の異なる部位がレジェンドビューアに表示される。例えば、図 6 に示されるグラフィカルビューアページ 600 では、開始座標 ウィンドウ 605 に 0 が入力されている点を除き、ページ 400 と同じように設定されている。このため、レジェンドビューア 612 に示されるコンティグ配列 602 および関連の座位 604 が、4467 塩基対分だけ上流に推移して、コンティグ SAU1c0039 の先頭にきている。図 4 に示されるコンティグ 414 の最下流 4467 に位置する塩基対は、ページ 600 のビューアではもはや見ることができない。また、ターゲットビューア 632 および配列深さビューア 652 には、対応する表示がなされている。

【0073】また、コンティグ ID ウィンドウ 407 は、ビューア 402 に表示されているコンティグを識別するのに加え、ユーザから異なるコンティグ ID の入力を受け付けることも可能である。異なるコンティグ ID の入力によって、新しいコンティグ ID に関連付けられたコンティグ配列のうちデフォルト数分の塩基対（座標 0 からスタートすることが好ましい）が、ビューアに関連付けられたデータベースからロードされて表示される。図 7 は、例えばコンティグ ID ウィンドウ 707 にコンティグ ID : SAU1c0016 を入力した際のグラフィカルビューアページ 700 を示している。このとき、レジェンドビューア 712 には、コンティグ SAU1c0016 のコンティグ配列 702 およびその関連座位 704 が表示されている。また、ターゲットビューア 732 および配列深さビューア 752 には、対応する表示がなされている。

●【0074】また、上述したように、グラフィカルビューア 400 は、種々のコマンドおよび制御機能を一覧表示するプルダウンメニューにアクセスするためにメニューバー 404 を備える。「File」プルダウンメニューには、保存や印刷等の、アプリケーションソフトウェアパッケージに見られる標準的なコマンドが一覧表示されている。「Edit」プルダウンメニューは、コンティグ配列長さのデフォルト表示数や、ビューア内で種々の特徴を表示するのに使用される色などの、グラフィカルビューアのパラメータを編集するためのカテゴリリストを提供する。

【0075】特に重要なのは、「View」プルダウンメニューである。この「View」プルダウンメニューは、ユーザが種々のビューア表示に含ませる機能を選択できるようにするとともに、座標上の配列表示 804 オプションをも備えている。図 8 A は、「View」プルダウンメニューが選択された状態のグラフィカルビューアページ 800 を示したものである。メニュー 802 において、座標上の配列表示 804 を選択すると、グラフィカルビューア 402 に示されたコンティグを構成するのに使用される配列が、その各配列のカバレッジを示す

座標とともに一覧表示されたページ 810 にアクセスすることができる。このリストから配列を 1 つ（例えば 2 番目の 806503054FI (5201, 5690) 812）選択し、配列データベースボタン 814 をクリックすると、グラフィックビューアに関連付けられたデータベースシステムであってコンティグを構成するのに使用される、未加工の配列データベースにアクセスする。このとき、図 8 C に示されるような、配列検索結果ページ 820 が返される。配列検索結果ページ 820 は、図 8 B で選択された配列 812 の実際のヌクレオチド配列 822 を示す。

【0076】図 9 は、本発明の好ましい一実施形態におけるグラフィカルビューアシステムが、遺伝子座位情報をユーザに返す際のプロセスを一般化して示すフローチャートである。このフローチャートでは、本発明の一実施形態における配列データをグラフィック表示するためのオプションをフローチャートの形式で示すために、本発明の好ましい一実施形態における主な機能のみが選択して示されている。よってこのフローチャートは、本発明を包括的に示すことを意図したものではない。

【0077】プロセス 900 がステップ 901 でスタートすると、ステップ 902 において、選択された座位に対応するデータおよびその関連のコンティグがグラフィカルビューアにロードされる。上述したように、座位は、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベース（この場合は細菌ゲノムデータベース）のユーザインタフェースの一部として提供される、HTML ページ内のリストから選択することができる。ステップ 904 では、コンティグ上の選択された座位がグラフィック表示される。このグラフィック表示は、選択座位に関連する生体分子配列データを異なる種々の側面から表示するための複数のコンポーネントを有することが好ましい。図 4 A および 4 B に示された特に好ましい実施形態では、レジェンドビューアと、ターゲットビューアと、配列深さビューアとの 3 つのコンポーネントを有するビューアによってグラフィック表示される。

【0078】さらなる入力またはズーム調整がなされなかった場合、上記プロセスは、ステップ 904 におけるグラフィック表示に続いてステップ 940 で終了する。しかしながら、ユーザがグラフィカルビューアを使用して選択座位またはその他座位に関する付加情報を抽出および表示したい場合には、ビューアは、その目的に合わせて付加的な機能を提供することができる。

【0079】グラフィカルビューアによって表示されたデータのグラフィック表示は、種々の方法で変更することが可能である。また、ビューア内の種々のオブジェクト（すなわち座位）を選択することによって、付加的な情報にアクセスすることも可能である。例えば、ユーザは、グラフィカルビューアページ内に設けられた一領域、例えば図 4 A のウィンドウ 407 に、新しいコンテ

ィグIDを入力しても良い。すると、判断ステップ906が肯定応答し、ステップ914において、ビューアに新しいコンティグおよびその座位がグラフィック表示される。また、ユーザは、例えば図4Aの開始座標ウィンドウ405に、新しい開始座標を入力しても良い。すると、判断ステップ908が肯定応答し、ステップ916において、グラフィック表示が調整されて新しい座標範囲のコンティグが表示される。上述したように、ユーザはさらに、グラフィック表示されたコンティグの特定部位に焦点を合わせることも可能である。このとき、判断ステップ910が肯定応答し、ステップ918において、本実施形態ではターゲットビューアのグラフィック表示が調整されて、そのコンティグが拡大表示される。上述したどの判断ステップにおいても、否定応答された場合にはグラフィカルビューアの表示は変化しないまま維持される。

【0080】上述したアクションの後かまたはそれに替わり、ユーザは、さらなる情報を得る目的でオブジェクトを選択することも可能である。好ましい一実施形態において、グラフィカルビューアのターゲットビューアコンポーネントに表示された座位は、その表示をクリックすることによって選択することが可能である。すると、判断ステップ920が肯定応答し、ターゲットビューアにおけるその座位の表示がカラーボックスで強調される。選択された座位に関して詳細な情報を得たい場合は、ユーザは、その座位の表示をダブルクリックする。すると、判断ステップ921が肯定応答し、ステップ922において、選択座位に関する詳細な情報を示したJava（登録商標）ページが表示される。

■【0081】本発明の好ましい一実施形態における他の特徴は、上述したように、グラフィカルアライメントビューアを備えることである。ユーザは、対象座位のアミノ酸配列の配置をいくつかのホモログス配列に対して示すグラフィカルビューアの表示を選択することが可能である。すると、判断ステップ924が肯定応答し、ステップ926において、そのアライメントがグラフィカルビューアにグラフィック表示される。

【0082】ユーザはまた、対象座位のホモログス座位およびパラログス座位のグラフィック表示を見るために、マルチプル生体ビューアの表示を選択することも可能である。例えば、判断ステップ920が肯定応答し、さらに判断ステップ928が肯定応答すると、ステップ930において、マルチプル生体ビューアにアクセスすることが可能となる。本発明の好ましい一実施形態におけるマルチプル生体ビューアの操作に関しては、図10A～10Eおよび図11とを参照しつつ、以下でさらに詳述する。

【0083】もちろん、選択されたオブジェクトの詳細を表示する選択判断と、マルチプル生体ビューアを表示する選択判断とは互いに独立しており、図9以外の方法

でも容易に可能である。なお、このシステムでは、ユーザがグラフィカルビューアモードをいつでも終了できるようにになっているが、この選択肢は図9には示されていない。

【0084】本発明のグラフィカルビューアからさらなる情報にアクセスするために使用可能なさらなる選択肢の1つに、対象座位とそのコンティグとに関連付けられた選択配列の実際のヌクレオチドまたはアミノ酸配列を表示する選択肢がある。好ましい一実施形態において、ユーザは、図4Aに示すようなグラフィカルビューアページのボタンをクリックすることによって、この選択肢を選択することが可能である。このとき、判断ステップ912が肯定応答し、ステップ932において配列のリスト（配列の識別子、および、ビューアに表示されたコンティグを構成する配列の座標）が表示される。そして、ユーザがそのリストから配列を1つ選択すると、ステップ932が肯定応答し、選択された配列の実際の（この場合は）ヌクレオチド配列が表示される。そして、ステップ940でプロセスが終了する。

【0085】本発明のグラフィカルビューアに表示される他のデータと同様に、本発明のこの態様で使用されるデータも、関連の生体分子配列データベースおよびシステムから得られる。このようなシステムの編成および操作は多様であり、本明細書で前掲したインサイト製薬の特許にその例が記述されている。当業者ならば、本明細書に記載された機能および表示の説明をもとにして、このようなシステムに本発明によるグラフィカルビューアを実装することが可能である。

【0086】マルチプル生体ビューア：上述したように、本発明のグラフィカルビューアを用いれば、1つまたはそれ以上の生体のゲノムに含まれる複数の部位を、対象座位ならびにそれに対応するパラログ（同じ生体のゲノムの他部位から得られる関連の座位）およびホモログ（別の生体のゲノムから得られる関連の座位）に基づいて、グラフィック表示および比較することが可能である。このようなマルチプル生体ビューアの好ましい一実施形態を、図10A～10Dを参照しつつ説明する。

【0087】図10Aは、図4Aおよび4Bに示されるような、メインのグラフィカルビューアページ1000を示したものである。図10Aでは、ボックス（ラバーバンド）1002が、グラフィカルビューア1010のレジェンドビューア1008コンポーネントにおいて表示されているコンティグ1006部位の一領域1004を取り囲むように配置されている。コンティグ1006のこの領域1004は、グラフィカルビューア1010のターゲットビューア1020コンポーネントに表示されており、そのカバレッジは、配列深さビューア1030コンポーネントに表示されている。ターゲットビューア1020において、座位SAU100242を囲むボックス1022は、その座位が選択されたことを示している。前

述したように、メインビューアページ 1000 は、マルチプル生体ボタン 1001 を備えている。

【0088】ターゲットビューアで座位が選択された際に、マルチプル生体ボタン 1001 をクリックすると、ビューアと関連付けられたデータベースが検索され、選択座位のホモログおよびパラログを含む全ライブラリが一覧表示される。図 10B は、図 10A のページで選択された座位 SAU100242 をもとに検索されたライブラリリストを示すウィンドウ 1040 である。ホモログおよびパラログの個々のリストにアクセスするためには、ユーザは、このウィンドウ 1040 に表示されたリストから 1 つまたはそれ以上のライブラリを選択すれば良い。そして、さらにマルチプル生体ボタン 1042 をクリックすると、個々のホモログおよび/またはパラログが検索されて表示される。図 10C は、図 10B の画面 1040 において選択されたライブラリから得られる座位 SAU100242 のホモログおよびパラログのリストを示すウィンドウ 1045 の一例を示したものである。

【0089】次に、ユーザは、初めに選択した座位（例えば SAU100242）、ならびに図 10C のリストに示される選択されたホモログス座位およびパラログス座位を、グラフィック表示するよう選択することが可能である。ウィンドウ 1045 のマルチプル生体ボタン 1046 をクリックすると、対象座位、ならびにそのホモログおよびパラログが、本発明の好ましい一実施形態におけるマルチプル生体ビューアにロードされて、対象座位、ならびに選択されたホモログおよびパラログが表示される。図 10D は、このようなマルチプル生体ビューアページ 1050 の一例を示したものである。

【0090】マルチプル生体ビューアページ 1050 は、コンティグ（SAU1c0039）上の選択された対象座位（SAU100242）、ならびにコンティグ上の選択されたホモログス座位およびパラログス座位をグラフィック表示する 1 つのパネルで構成されたマルチプル生体ビューア 1052 を提供する。図 10D に示すビューア 1052 は、1 つのページに 5 つのコンティグ：SAU1c0039 1061、PRT1c0129 1062、SAU2c0391 1063、SEP1c0220 1064、SHA1c0122 1065 をグラフィック表示している。コンティグ 1061 は、その座位とともに示されており、より明確に識別できるように太字イタリック体で表示された選択座位 SAU100242 1071 を含んでいる。図 10D に示されている一実施形態では、座位を、座位 ID よりもむしろヒットに関する記述によって注釈している。他の各コンティグもまた、横に沿って示された座位、ならびに太字イタリック体で示された SAU100242 にホモログス座位（座位 1072、1073、1074、および 1075）を伴って表示されている。

【0091】マルチプル生体ビューア 1052 は、本発明のグラフィカルビューアによる生体分子配列情報の効率的な伝送を示す別の例である。上述したように、太字

のイタリック体かまたは特定の色を使用する等の他の方法で、選択座位ならびにそのホモログ座位およびパラログ座位を表示することによって、対象座位として区別している。図 10D に示されるように、グラフィック表示されたコンティグの対象座位がページ 1052 に配置されているため、種々のコンティグ上の近くの座位を、視覚的に容易に比較することができる。この視覚的な表示は、図 10E を参照にしつつ以下で説明する補体機能を使用することにより、さらに向上させることが可能である。

【0092】本発明のこの実施形態におけるそのようなグラフィカルビューアの機能を利用するには、マルチプル生体ビューアページ 1050 に設けられたプルダウンメニュー選択 1053 をクリックすれば良い。このメニュー選択は、図 4A で説明したような機能を提供する「File」、「View」、「Help」選択を含んでいる。「Show」選択 1054 からは、図 10C に示すウィンドウに一覧表示されてマルチプル生体ビューアにロードされた全座位のリストにアクセスすることができる。「Show」プルダウンメニューから座位を 1 つ選択することにより、ユーザは、その座位を、それが属するコンティグとともに表示するか隠すかを決定することができる。座位をクリックすることにより、ユーザは、その座位を表示するか隠すかを決定することができる。「Show」メニューを使用して、コンティグに関して同様な決定を行うことも可能である。

【0093】補体メニュー選択 1055 により、ユーザは、データからの目立った情報の抽出を容易にするために、コンティグおよび座位のグラフィック表示を操作することができる。特に、補体メニュー選択 1055 により、ユーザは、マルチプル生体ビューア 1052 に表示された任意のコンティグ上で、逆補体を実施することが可能となる。こうすれば、ビューア 1052 に表示されたホモログス座位およびパラログス座位が同じ読み方向で表示されるため、ユーザは、対象座位の近くの関連座位のパターンをより容易に見つけることができる。図 10E に示されるマルチプル生体ビューアページ 1080 では、ページ 1050 に示された対象座位を、補体機能の使用によって同じ読み方向で示すことにより、コンティグ 1062、1063、1065 の逆補体を表示している。

【0094】本明細書で説明する他の機能と同様、この補体機能へのショートカットも、当業者に周知の方法によって使用可能とすることができる。例えば、座位（コンティグ）の補体を表示するには、グラフィカルビューアが作動しているコンピュータシステムと接続されたキーボード上のシフトキーを押さえながら、そのコンティグをクリックすれば良い。

【0095】図 11 は、本発明の好ましい一実施形態におけるマルチプル生体ビューアの操作を一般化したプロ

10

20

30

40

50

セスを示すフローチャートである。プロセス 1100 はステップ 1101 でスタートし、ステップ 1102 では、マルチプル生体ビューアシステムが、例えば図 10A のターゲットビューア内の座位のクリックに伴い、対象座位の選択を受け入れる。ステップ 1104 では、選択された対象座位に対するホモログス座位またはパラログス座位を含むライブラリのリストが、ウィンドウに表示される。この表示は、例えば図 10A のマルチプル生体ボタン等のボタンをユーザがクリックすることによって始まる。次に、ステップ 1106 において、システムは、リストから 1 つまたはそれ以上のライブラリの選択を受け取る。ステップ 1108 では、選択されたライブラリから得られるリスト、すなわち、選択された対象座位に対するホモログス座位またはパラログス座位のリストが、ウィンドウに表示される。ステップ 1110 において、システムは、表示されたリストから、選択された対象座位に対するホモログス座位またはパラログス座位の選択を受け取る。次いでステップ 1112 において、選択座位およびその関連のコンティグがマルチプル生体グラフィカルビューアにグラフィック表示される。好ましい一実施形態において、ビューアは、グラフィック表示されたデータの比較を容易にするために、全てのコンティグおよび座位を 1 つのパネル内に表示する。プロセスはステップ 1114 で終了する。

【0096】実施：本発明は、システムまたは方法として実施でき、さらに、本明細書に記述した種々の操作を実行するためのプログラム命令等を含んだ種々のコンピュータ読み取り可能な媒体に組み込むことができる。上述したように、上記のシステムは、生体分子配列リレーショナルデータベースシステムとの関連のもとで実施されることが好ましい。上記の方法は、コンピュータで実行される方法であり、一般に、そのようなシステムの操作を含む。上記の媒体は、任意のコンピュータ読み取り可能な媒体で良い。コンピュータ読み取り可能な媒体の例としては、ハードディスク、フロッピーディスク、磁気テープ等の磁気媒体、CD-ROM ディスク等の光媒体、プロプティカルディスク等の光磁気媒体、ならびに読み取り専用メモリ (ROM) およびランダムアクセスメモリ (RAM) 等の、プログラム命令を格納および実行するために構成されたハードウェアデバイス等が挙げられるが、これらに限定はされない。本発明はまた、エアウェーブ、光路、電線路等の適切な媒体を通じて伝わるような、搬送波に組み込まれていても良い。

【0097】結論：以上では、理解を明確にする目的で本発明を細部にわたって説明したが、添付した特許請求の範囲の範囲内であれば、一定の変更および修正を加えられることは明らかである。また、本発明による方法、媒体、およびシステムを実施する代替の方法が、数多く存在することを留意する必要がある。上述したように、本発明の範囲は、本発明との関連のもとで説明した、細

菌ゲノムデータベースシステムにおける使用に限定されるものではない。当業者ならば、種々のコンピュータベースの生体分子配列データベースシステムとの関連のもとで本発明を使用する方法を、本明細書でなされた説明から理解することが可能である。例えば、本発明のグラフィカルビューアを、他のタイプおよび形式の核酸配列の格納および分析や、発現核酸もしくはアミノ酸の配列の格納および分析に利用されるデータベースシステムとの関連のもとで使用することも可能である。本実施形態は例示であり、また、非限定的なものであり、本発明は、本明細書で述べた細目に限定されず、添付した特許請求の範囲の範囲および均等物の範囲内で変更することが可能である。

【図面の簡単な説明】

【図 1 A】本発明の一実施形態に従ってデータベースサービスを提供するためのクライアントサーバインタラネットを示すブロック図である。

【図 1 B】ユーザからの照会に応じて生物学的情報を提供するために、図 1 A のクライアントサーバインタラネットで利用される種々のソフトウェア文書およびエンティティを示す概略図である。

【図 1 C】本発明の好ましい一実施形態に従ったグラフィカルビューア機能のアクセス可能性を、生体分子配列データベースに関連させて示すブロック図である。

【図 2】本発明の一実施形態に従った生体分子配列グラフィカルビューアで表示される座位の選択に適した、ゲノム配列データベースのグラフィカルユーザインタフェースとしてのコンティグ結果ページを示す画面 (HTML ページ) である。

【図 3】本発明の一実施形態に従った生体分子配列グラフィカルビューアにアクセスするのに適した、ゲノム配列データベースのグラフィカルユーザインタフェースとしての座位情報ページを示す画面である。

【図 4 A】本発明の一実施形態に従った生体分子配列グラフィカルビューアのメインページを示す画面である。

【図 4 B】ズーム機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示す画面である。

【図 5 A】本発明の一実施形態に従った選択オブジェクト詳細ウィンドウである。

【図 5 B】本発明の一実施形態に従ったアライメントビューアを示す画面である。

【図 5 C】本発明の一実施形態に従ったアライメントビューアを示す画面である。

【図 6】新しい開始座標機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示した画面である。

【図 7】新しいコンティグ ID 機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示した画面であ

る。

【図 8 A】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図 8 B】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図 8 C】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図 9】本発明の一実施形態に従った生体分子配列グラフィカルビューアで遺伝子座位情報を見るためのプロセスフローを示したフローチャートである。

【図 10 A】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図 1 A】

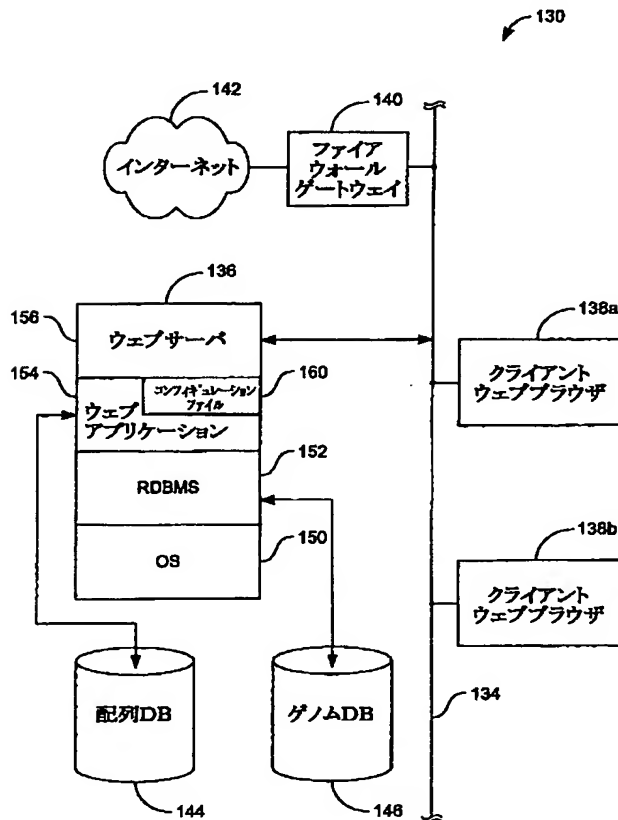


FIG. 1A

【図 10 B】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図 10 C】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図 10 D】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

10 【図 10 E】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図 11】本発明の一実施形態に従った生体分子配列グラフィカルビューアでマルチプル生体の遺伝子座位情報を見るためのプロセスフローを示したフローチャートである。

【図 1 B】

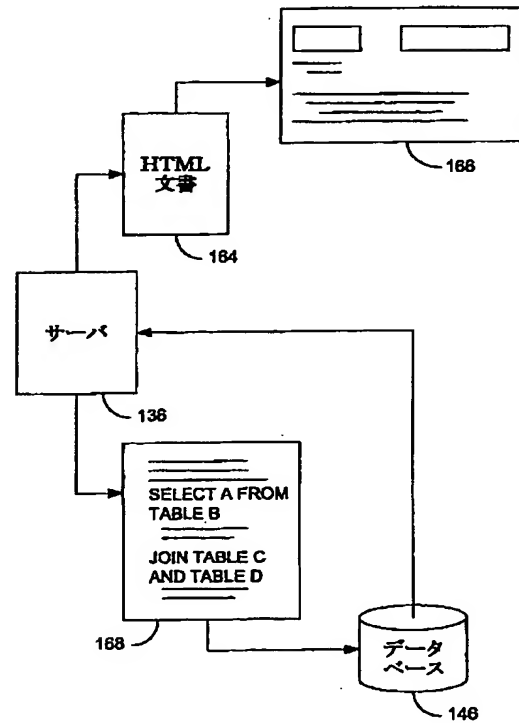


FIG. 1B

【図 1 C】

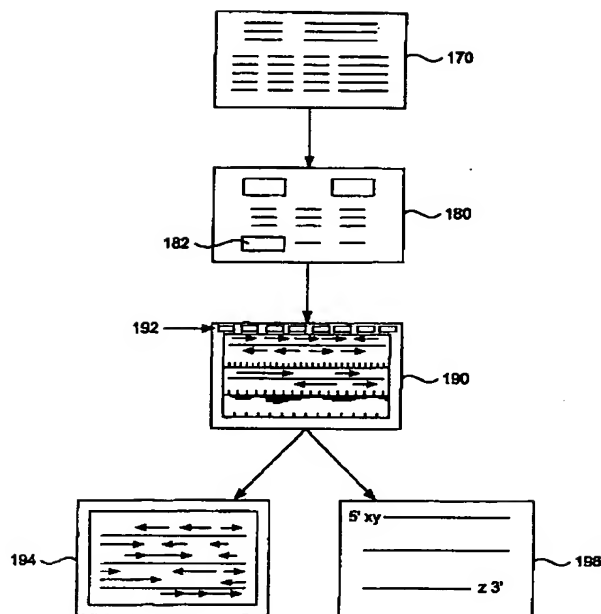


FIG. 1C

【図 8 B】

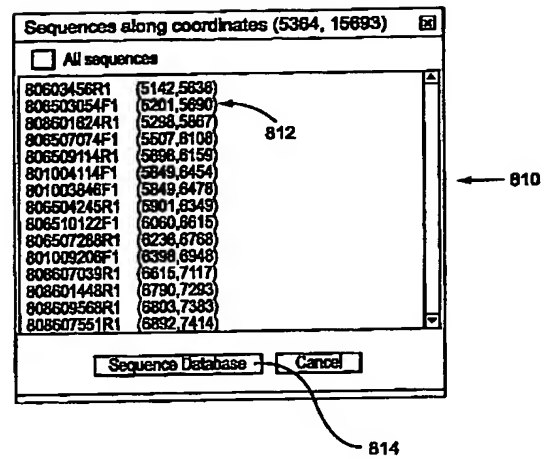


FIG. 8B

【図 10 B】

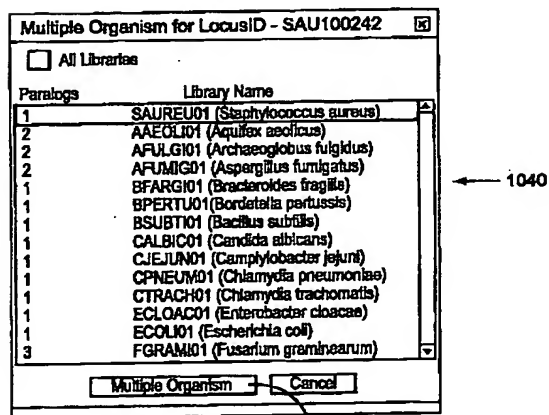


FIG. 10B

【図 10 C】

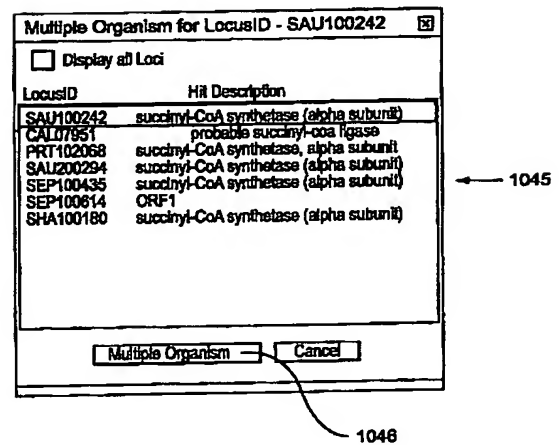


FIG. 10C

【図2】

Contig Results - Netscape									
Contig Results									
Main Menu	Org Info	Genome	Contig Summary	Contig	Locus Info	Proteins	Comparisons	Sequences	Help
Library : SAUREU01 Staphylococcus aureus Contigs: 48 Contigs ORFs: 134									
Contig: SAU1c0039 Length: 138277 bp Segs: 1846 Base Pairs Selected: 1 to 138277									
LocusID	Hit ID	Hit Description	Hit Organism	E-Value	Base Pairs	Llrs (48)	Seqs		
SAU102730	g2633184	yfh0	Bacillus subtil	3.4e-51	22-2168 (-)	8	26		
SAU103442	LUR			1	2170-2722 (-)	0	22		
SAU100689	g2633977	similar to hypothetical proteins	Bacillus subtil	4.8e-62	2723-3559 (+)	24	30		
SAU102070	g2633978	ribonuclease H	Bacillus subtil	3.2e-39	3594-4352 (+)	34	9		
SAU100241	g2462967	putative succinyl-coA synthetase beta ch	Bacillus subtil	6.8e-123	4487-5621 (+)	35	13		
SAU100242	g2633982	succinyl-CoA synthetase (alpha subunit)	Bacillus subtil	2.9e-97	5655-6542 (+)	35	10		
SAU102767	g3767593	LytN	Staphylococcus	4.8e-23	6849-8998 (+)	3	7		
SAU101151	g3767595	Eprh	Staphylococcus	0	7029-8270 (+)	13	18		
SAU101152	g2633984	ORF4	Staphylococcus	3.5e-109	8514-9314 (+)	36	13		
SAU101153	g2462971	DNA Topoisomerase I	Bacillus subtil	0	9497-11563 (+)	48	32		
SAU101154	g2633986	Gid protein	Bacillus subtil	0	11728-13026 (+)	43	36		
SAU101155	g2633987	Integrase/recombinase	Bacillus subtil	7.1e-51	13440-14333 (+)	38	27		
SAU101156	g2633987	beta-type subunit of the 20S proteasome	Bacillus subtil	9.7e-52	14348-14881 (+)	21	17		
SAU101157	g2633988	ATP-dependent Clp protease-like	Bacillus subtil	7.4e-95	14971-16350 (+)	40	16		
SAU101158	g2633989	transcriptional regulator	Bacillus subtil	1.3e-82	16376-17139 (+)	9	8		
SAU100275	g2634021	ribosomal protein S2	Bacillus subtil	2.8e-73	17493-18200 (+)	47	12		

FIG. 2

200

[図3]

Locus Query - Netscape

Locus Information

Search by LocusID:
 Display: ☒ Locus Details
 ☐ Loci around the selected Locus

~ 302

Locus ID: SAU100241 Type: Homology Gene Category: Small molecule metabolism
 Contig ID: SAU10039 Position: 5 of 150
 Amino Acids: 385 Nucleotides: 1155 Seqs: 13
 Homologs: 46 Paralogs: 1 Llbs: 35

Top hits: FASTA of ORF against pagenpept110

Hit ID	Hit Description	Hit Organism	E-Value	ORF Coverage
g2462967	putative succinyl-coA synthetase beta ch	Bacillus subtil	<u>6.8e-123</u>	100%
g2633981	succinyl-CoA synthetase (beta subunit)	Bacillus subtil	<u>6.8e-123</u>	100%

300

FIG. 3

【図4A】

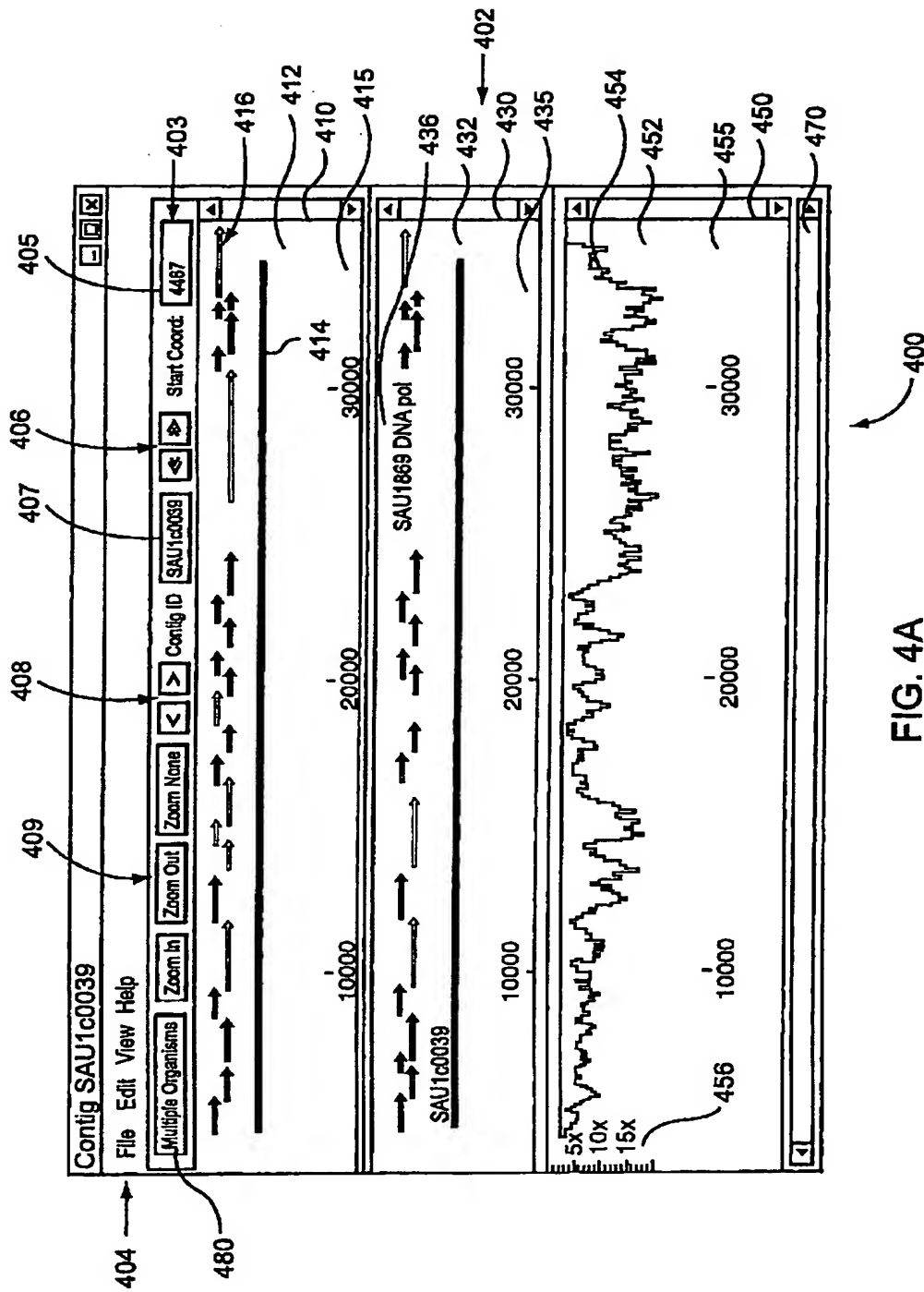
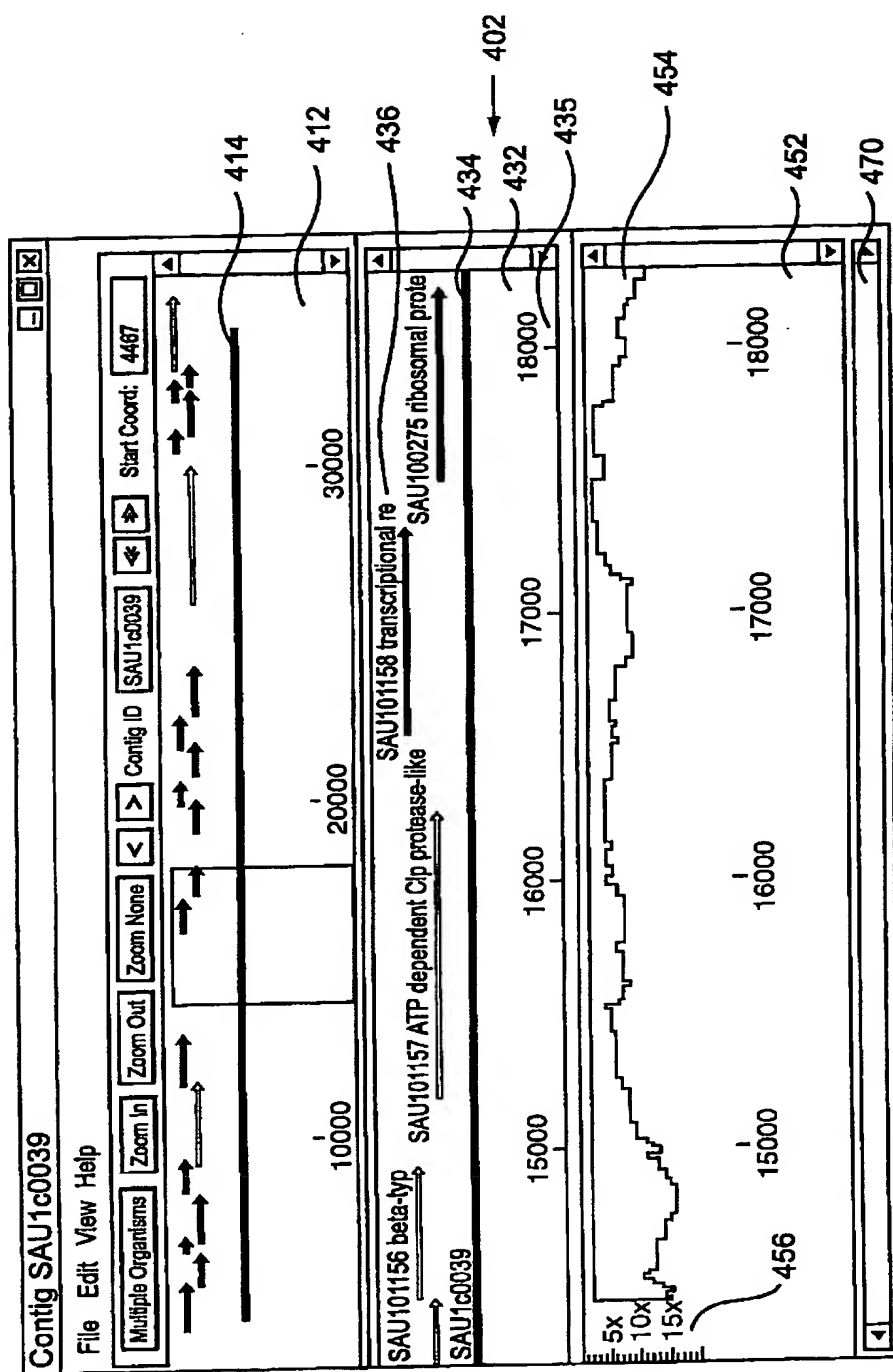


FIG. 4A

【図 4 B】



【図 5 A】

The image shows a software window titled "Selected Object Details" with a close button in the top right corner. Inside the window, there is a section titled "Details" which contains the following information:

- Locus ID: SAU101156
- Gene Category: Macromolecule metabolism
- Hit Description: beta-type subunit of the 20S proteasome
- EValue: 3e-68
- Libs: 22
- Seqs: 17
- Base Pair: 14348..14881(+)
- Hit Data Source: genpept111
- Locus Type: Homology Based ORF
- HitID1: g2633987
- HitID2: g535349
- HitID3: g488725
- HitID4: g1573473
- HitID5: g2984006

At the bottom of the window, there are two buttons: "Dismiss" and "Alignment".

500

502

FIG. 5A

【図 5 B】

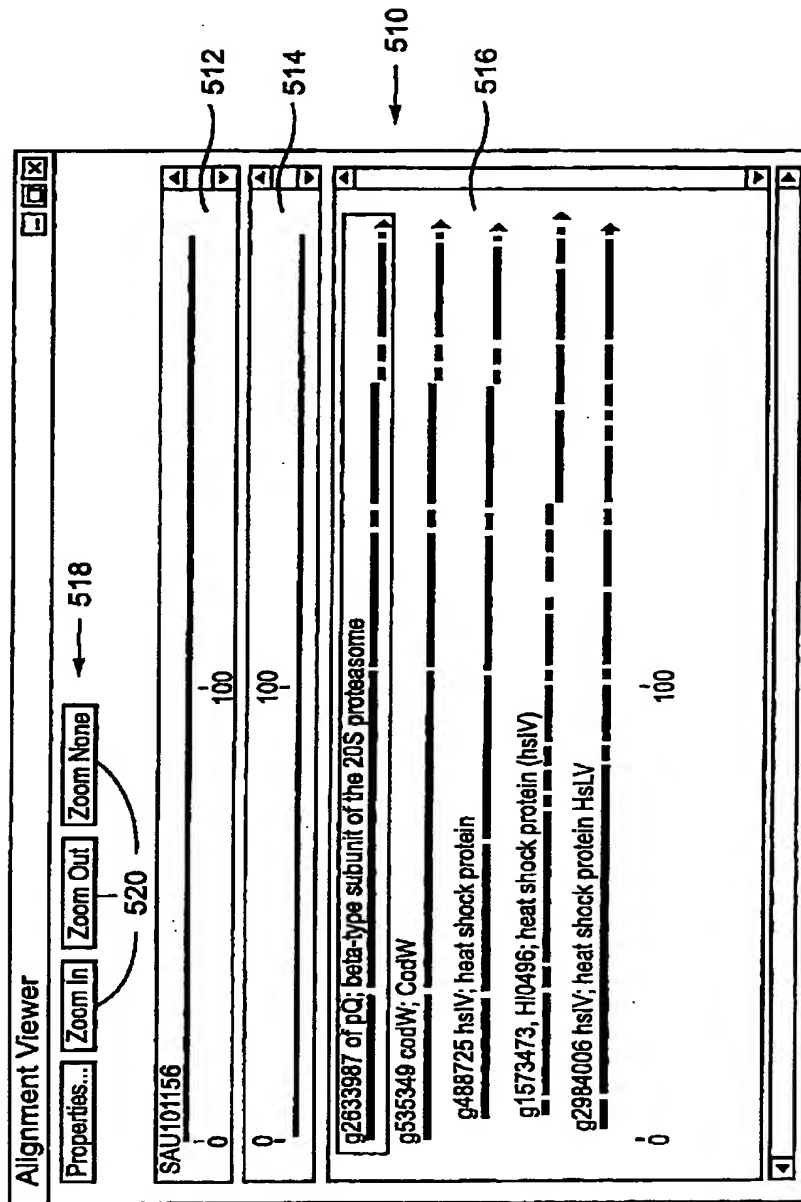


FIG. 5B

【図 5C】

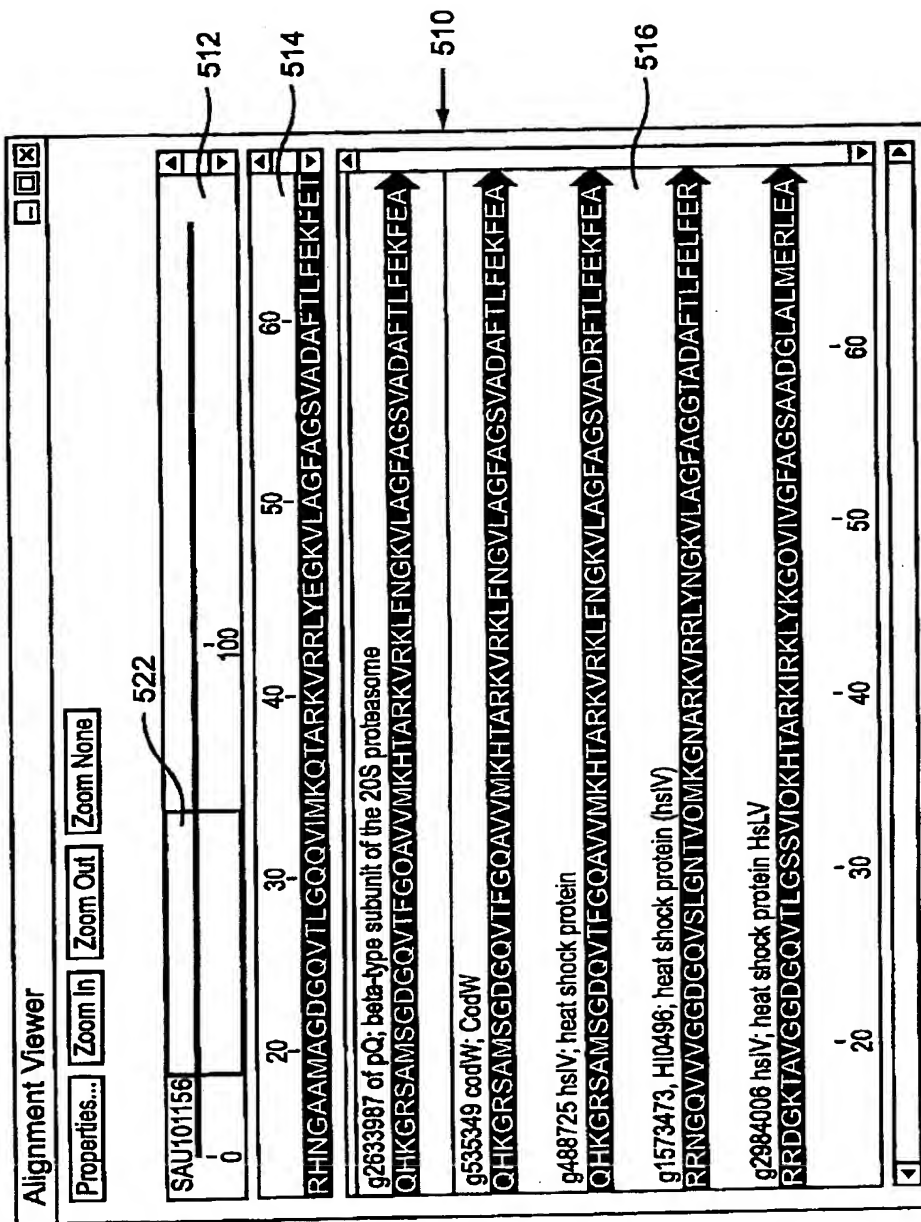


FIG. 5C

【図6】

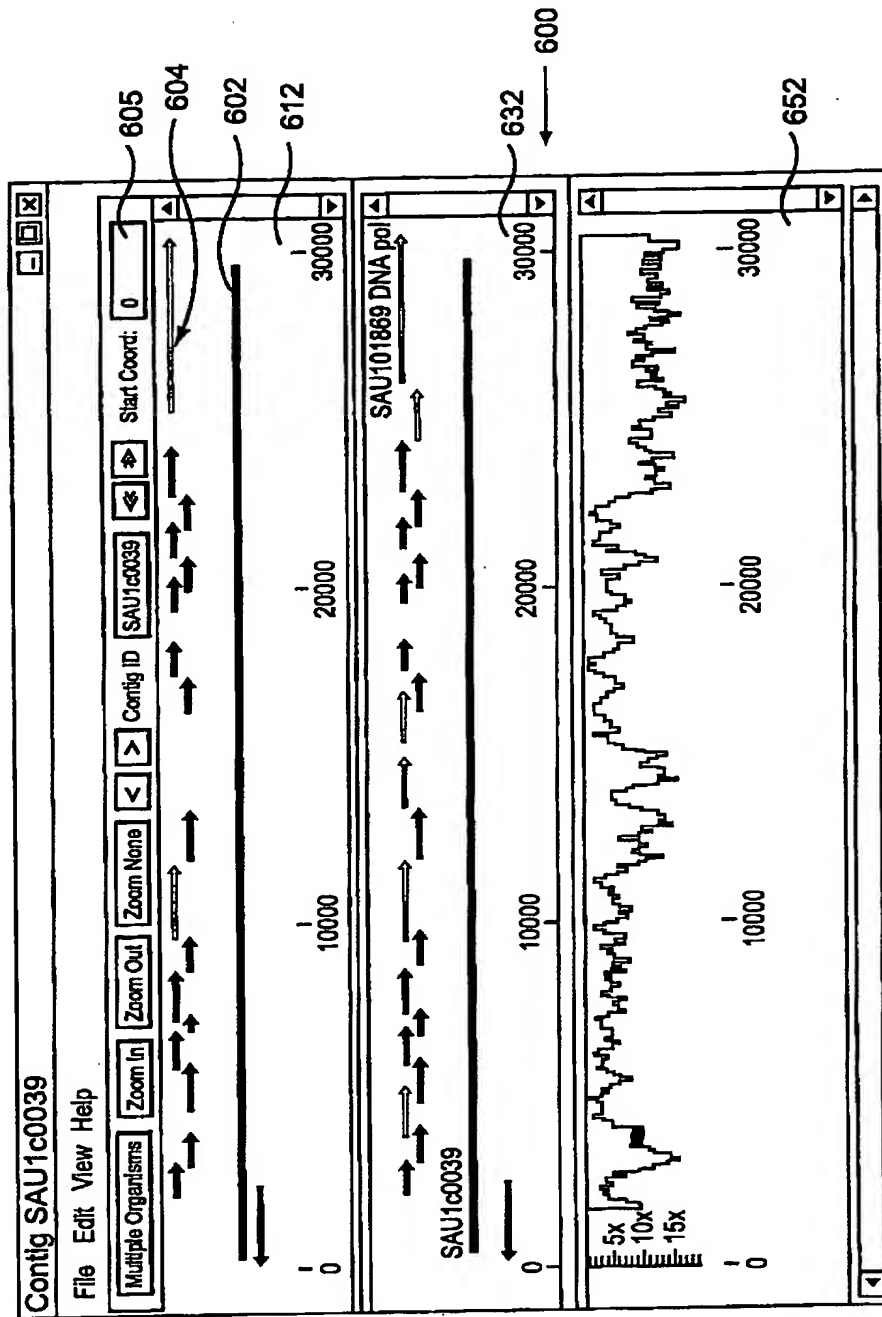


FIG. 6

【図7】

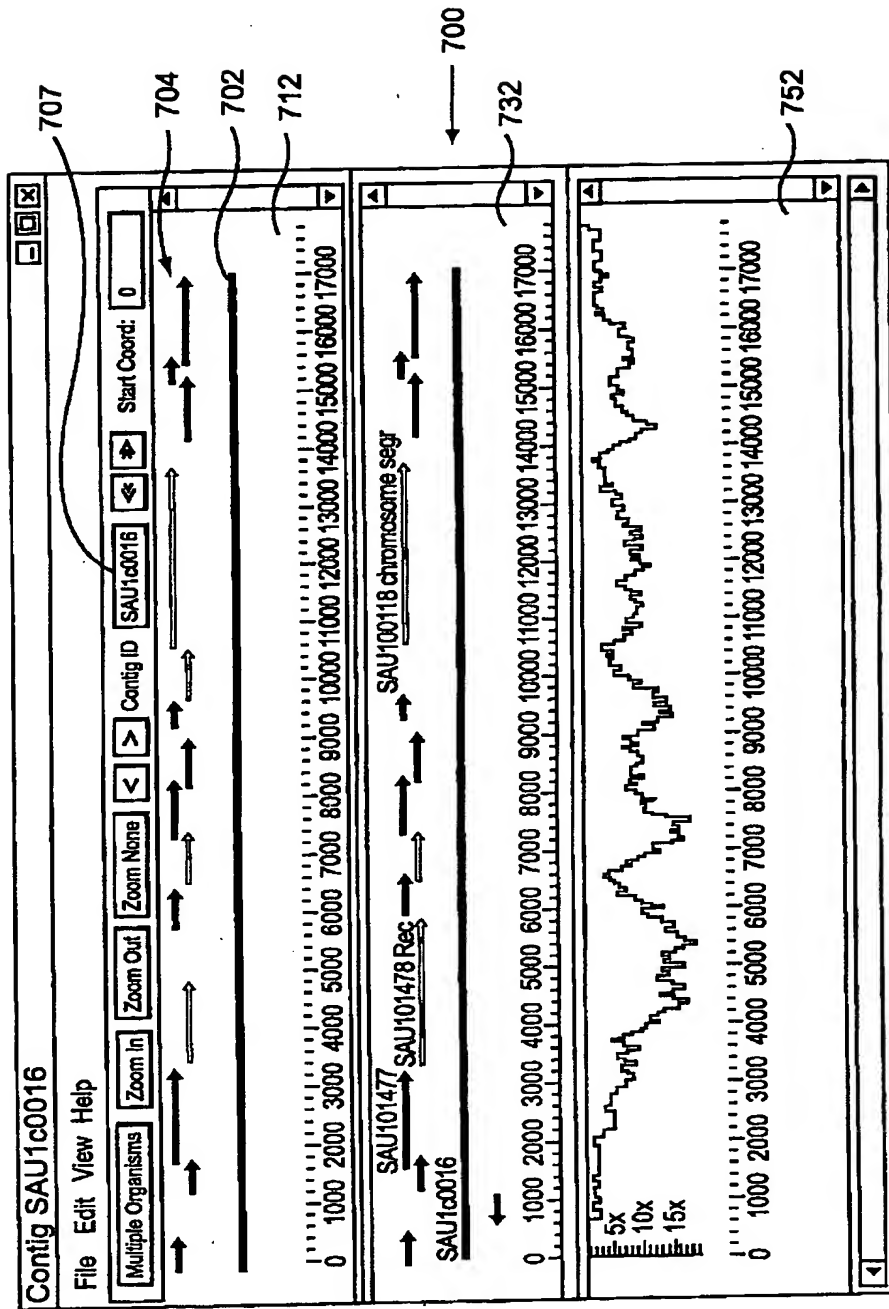


FIG. 7

【図 8 A】

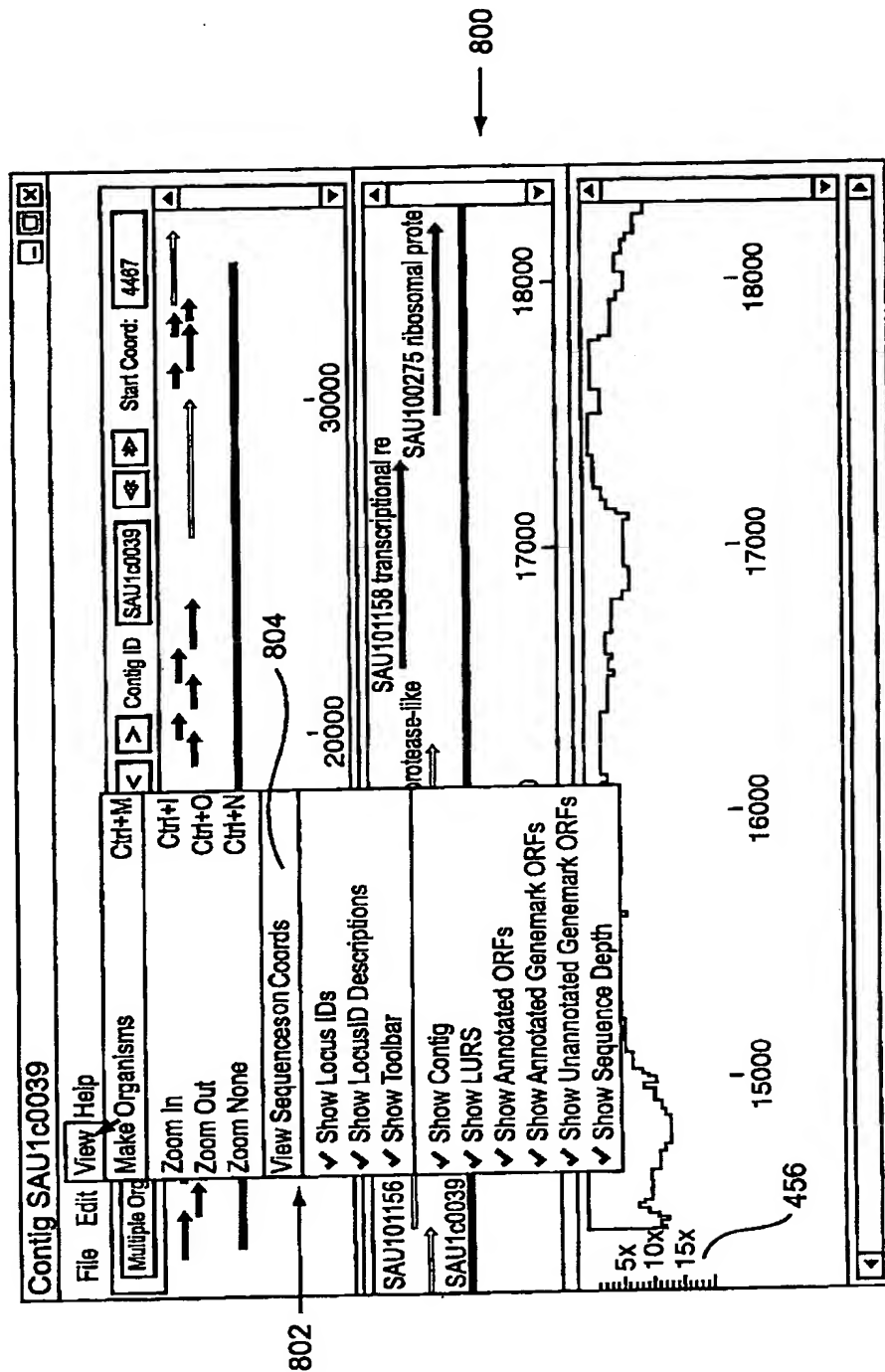


FIG. 8A

【図8C】

Sequence Retrieval Results - Netscape

Sequence Retrieval Results

☒ >806503054F1 saureu01

```

GTCTAAACATGATTATCATACATTGCATTAGTGGTACATCGGATGTATGGTGAATGGTGCAGGTTAGCCAT
GGCAACAATGGATACGAATTAAATCATTTTCGGTGGAAACCCAGCCAAATTTCTAGATGCAAGCGGAAAGCGCTACTAG
AGAAAAGTAACTGAAGCATTAAATCATTTTAGGTGATGAAATGTTAAAGGATTTTGTAAACATTTTCGG
TGGCATTGAATGTGATGTTATCGCAGAAGGTATCGTTGAAGCTGTAAGAAGTAGATTAACTTTACCACT
AGTTGTACGCTTAGAAGGTACAAATGTTGAGTTAGGTAAATAATCTTAAAGACTCAGGATTAGCAATTGAACC
AGCAGCAACATGGCTGAAGGTGCACAAAATTTGTTAACTAGTCAAGAAGCATAAGGAAAGGATGGGAGCAC
TAAGATGAGTGTATTTATAGATAAGAATACTAAAGTAAAT
  
```

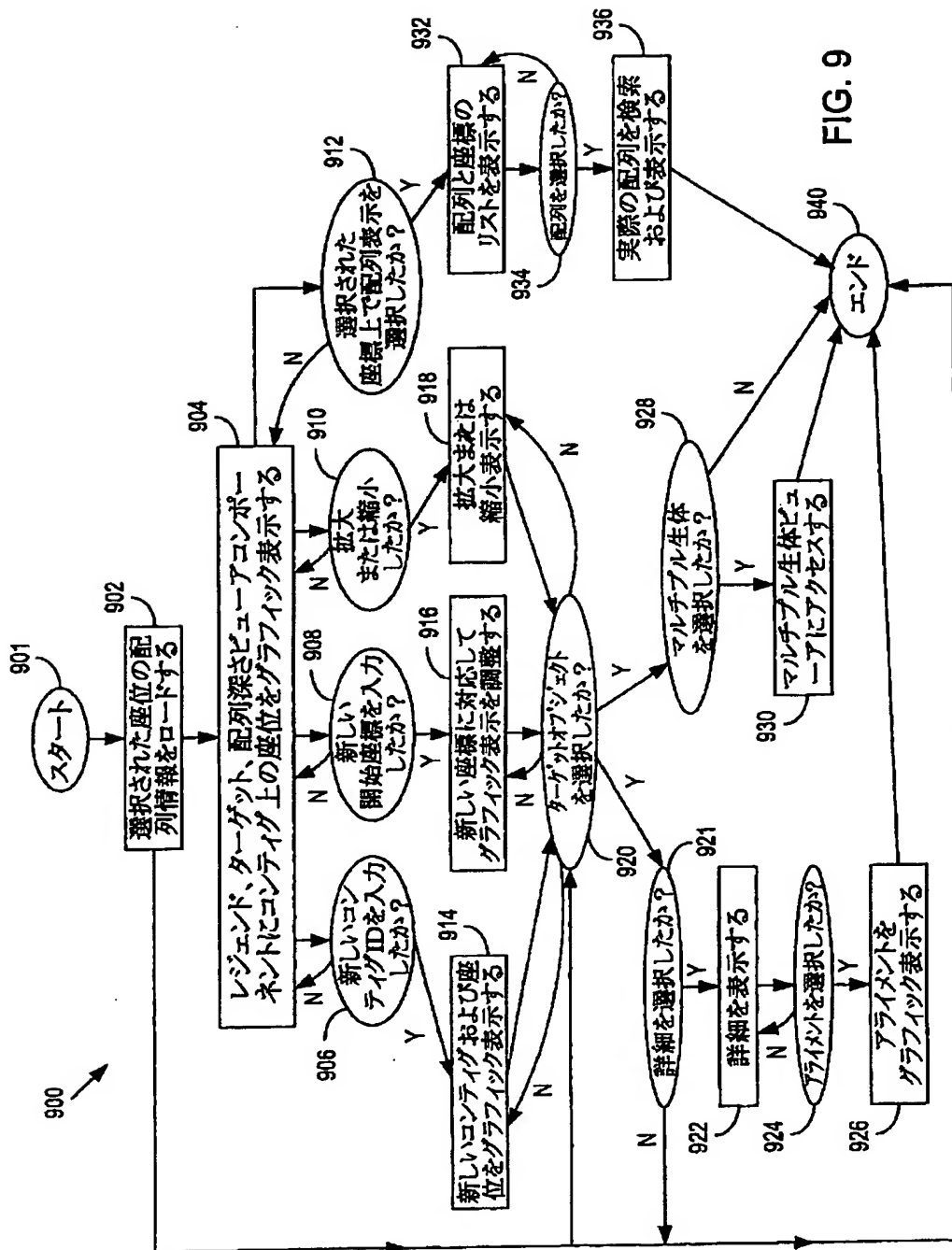
822

Submit sequences to:

FIG. 8C

820

【図9】



【図10A】

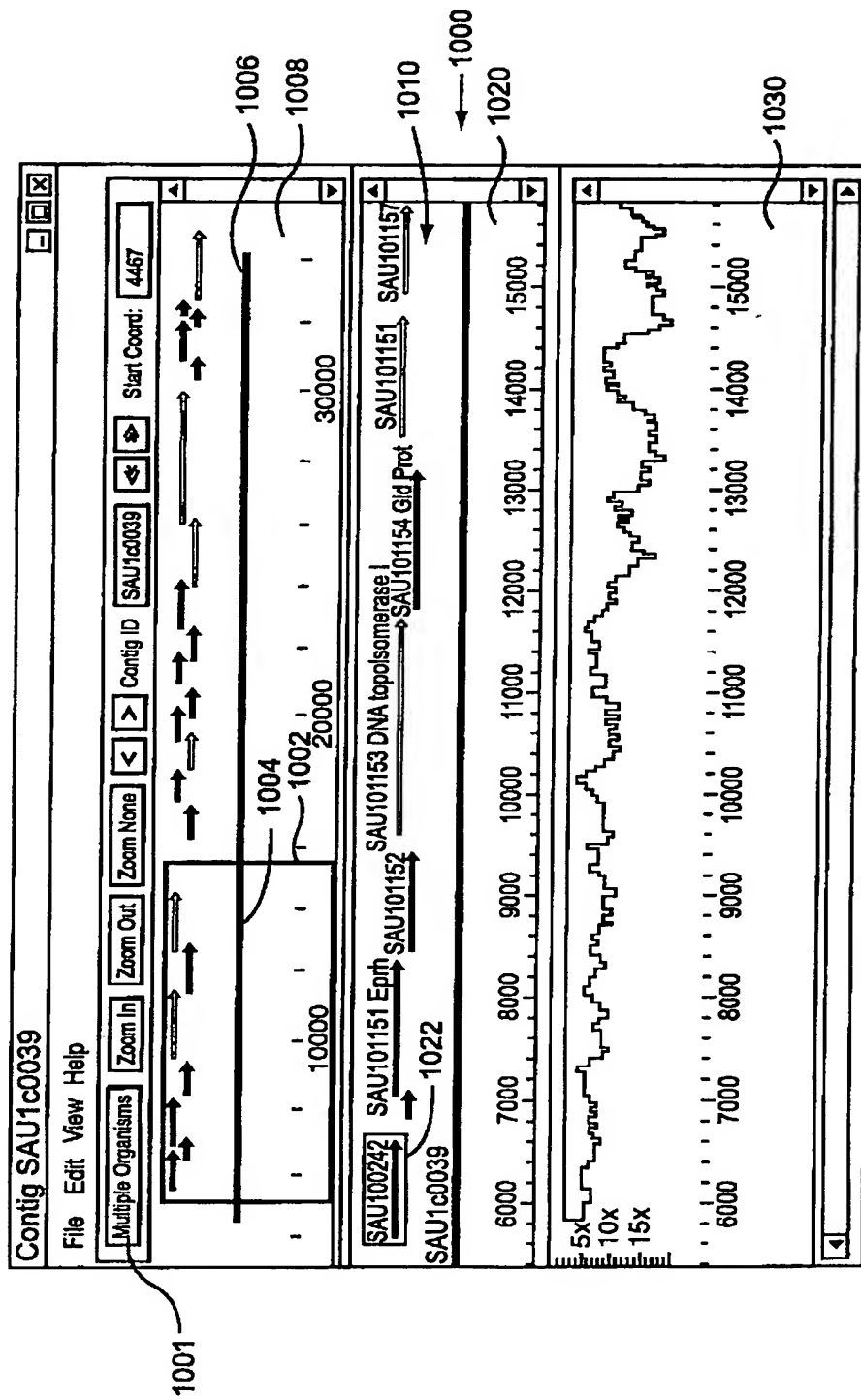


FIG. 10A

【図10D】

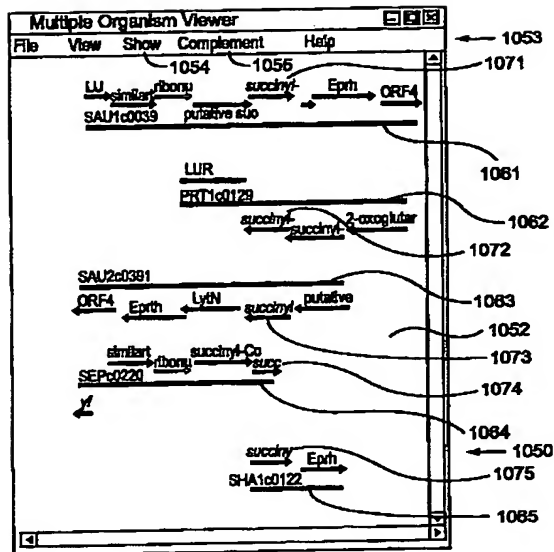


FIG. 10D

【図10E】

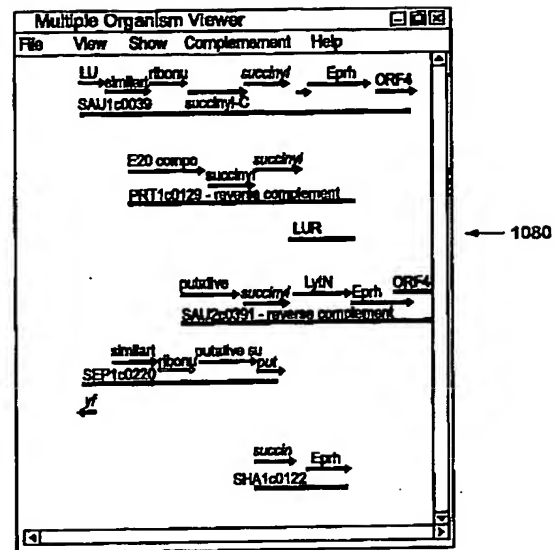


FIG. 10E

【図11】

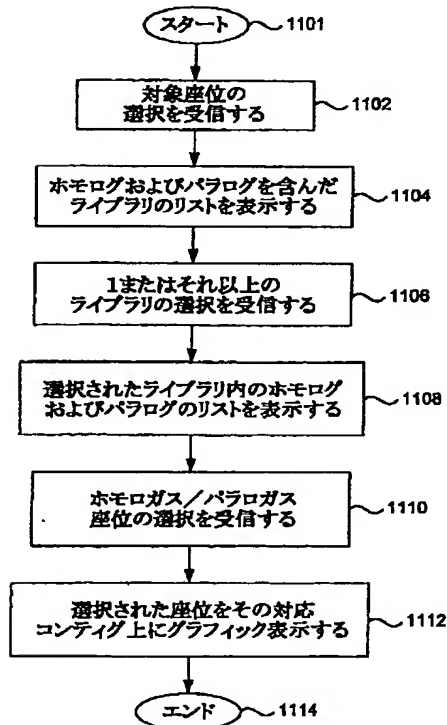


FIG. 11

フロントページの続き

(51) Int. Cl. ⁷	識別記号	F I	テ-マ-ド (参考)
G 0 6 F 3/00	6 5 5	G 0 6 F 3/00	6 5 5 B
// C 1 2 N 15/09		C 1 2 N 15/00	A
(72) 発明者 フランク・ディー・ルソー アメリカ合衆国 カリフォルニア州94086 サニーバイル, ロゼッタ・コート, 939		(72) 発明者 レイチェル・ジェイ・ライト アメリカ合衆国 カリフォルニア州94043 マウンテン・ビュー, アナ・アベニュー, 339	
(72) 発明者 ジョー・ドン・ヒース アメリカ合衆国 カリフォルニア州95087 サニーバイル, グレーブ・アベニュー, 856		(72) 発明者 ピーター・エイ・コヴィッツ アメリカ合衆国 カリフォルニア州94116 サン・フランシスコ, 29番・アベニュー, 2442	
(72) 発明者 ステファニー・エフ・ベリー アメリカ合衆国 カリフォルニア州95117 サン・ホセ, サニーゲート・コート, 3564		(72) 発明者 イヴォンヌ・アール・ボルド アメリカ合衆国 カリフォルニア州94087 サニーバイル, ジェイムスタウン・ドライブ, 1109	
		(72) 発明者 リー・クロフォード アメリカ合衆国 カリフォルニア州94118 サン・フランシスコ #204, サクラメント・ストリート, 3570	

【外国語明細書】

1. TITLE OF INVENTION

Graphical Viewer for Biomolecular Sequence Data

2. CLAIMS

1. A method implemented in a computer system for presenting biomolecular sequence data, comprising:

retrieving biomolecular sequence data from a database in response to a user query; and

graphically depicting elements of the biomolecular sequence data in a user interface for said computer system.

2. The method of claim 1, wherein said graphical depiction comprises a plurality of panels, at least one of said plurality of panels graphically depicting depth of coverage information for a portion of a biomolecular sequence depicted in at least one other of said plurality of panels.

3. The method of claim 2, wherein said plurality of panels are comprised within a single frame.

4. The method of claim 3, wherein said plurality of panels provide graphical depictions representing different aspects of said biomolecular sequence data.

5. The method of claim 4, wherein said biomolecular sequence data comprises gene locus data.

6. The method of claim 5, wherein said plurality of panels comprises three panels.

7. The method of claim 6, wherein said three panels comprise a first panel graphically depicting at least a portion of a contig and its associated loci, a second panel graphically depicting at least a portion of the contig depicted in said first panel and annotated loci associated with the portion, and a third panel graphically depicting depth of coverage information for the portion of the contig depicted in the second panel.

8. The method of claim 7, wherein said third panel graphically depicts depth of coverage information for the portion of the contig depicted in the second panel.

9. The method of claim 1, wherein said method is implemented in Java programming language.

10. A method implemented in a computer system for presenting biomolecular sequence data, comprising:
retrieving biomolecular sequence data for a plurality of homologous loci from a database in response to a user query; and
graphically depicting at least some of the homologous loci in a user interface for said computer system.

11. The method of claim 10, wherein said graphical depiction comprises a single panel.

12. A computer system, comprising:
a database including biomolecular sequence data;
a user interface capable of
receiving a query relating to the biomolecular sequence data, and
graphically displaying the results of said query.

13. The system of claim 12, wherein said graphical depiction comprises a plurality of panels, at least one of said plurality of panels graphically depicting depth of coverage information for a portion of a biomolecular sequence depicted in at least one other of said plurality of panels.

14. The system of claim 13, wherein said plurality of panels are comprised within a single frame.

15. The system of claim 14, wherein said plurality of panels provide graphical depictions representing different aspects of said biomolecular sequence data.

16. The system of claim 15, wherein said biomolecular sequence data c

omprises gene locus data.

17. The system of claim 16, wherein said gene locus data is depicted in three panels comprising a first panel graphically depicting at least a portion of a contig and its associated loci, a second panel graphically depicting at least a portion of the contig depicted in said first panel and annotated loci associated with the portion, and a third panel graphically depicting depth of coverage information for the portion of the contig depicted in the second panel.

18. A computer-readable medium containing programmed instructions arranged to graphically display biomolecular sequence data, the computer-readable medium including programmed instructions for:

retrieving biomolecular sequence data from a computer system database in response to a user query; and

graphically depicting elements of the biomolecular sequence data in a user interface for the computer system.

3. DETAILED DESCRIPTION OF INVENTION

BACKGROUND OF THE INVENTION

The present invention relates generally to the field of bioinformatics.

In particular, the invention relates to methods, media and systems for graphically displaying computer-based biomolecular sequence information.

Informatics is the study and application of computer and statistical techniques to the management of information. Bioinformatics includes the development of methods to search computer databases of biomolecular sequence information (e.g., nucleic acid and protein) quickly, to analyze and display biomolecular sequence information, and to predict protein sequence, structure and function from DNA sequence data.

Increasingly, molecular biology is shifting from the laboratory bench

to the computer desktop. Today's researchers require advanced quantitative analyses, database comparisons, and computational algorithms to explore the relationships between sequence and phenotype. Thus, by all accounts, researchers cannot and will not be able to avoid using computer resources to explore gene sequencing, gene expression, and molecular structure.

One use of bioinformatics involves studying an organism's genome to determine the sequence and placement of its genes and their relationship to other sequences and genes within the genome or to genes in other organisms. Such information is of significant interest in biomedical and pharmaceutical research, for instance to assist in the evaluation of drug efficacy and resistance. To make genomic information manipulation easy to perform and understand, sophisticated computer database systems have been developed. Incyte Pharmaceuticals, Inc. of Palo Alto, CA, has developed several such databases, including some in which genomic sequence data is electronically recorded and annotated with information available from public sequence databases. Examples of such public sequence databases include GenBank (NCBI) and SWISSPROT. The resulting information is stored in a relational database that may be employed to determine relationships between sequences and genes within and among genomes.

While genetic data processing and relational database systems such as those developed by Incyte Pharmaceuticals, Inc. provide great power and flexibility in analyzing genetic information, further improvements in these systems will help accelerate biological research for numerous applications.

One area of interest in this regard is the display of biomolecular sequence information. As noted above, an important goal of genome research is to determine the sequence and placement of an organism's genes and their relationship to other sequences and genes within the genome, to genes i

in other organisms, and to related protein sequences. The ability to clearly and effectively display gene loci information for a given organism or organisms would greatly assist this task.

Accordingly, the development of a display tool which allows a user to clearly and effectively display gene loci information for a given organism or organisms and/or other biomolecular sequence information is desirable.

SUMMARY OF THE INVENTION

The present invention meets this need by providing methods, media and systems for graphically displaying computer-based biomolecular sequence information. Generally, biomolecular sequence information may be graphically depicted in a variety of different forms in accordance with the present invention. The sequence information may be composed of nucleotide or amino acid sequence information or both. The graphical depictions may be in several different formats providing different information relating to the sequences, and may be displayed in one or more screens of a computer user interface.

A graphical viewer in accordance with the present invention preferably has a plurality of panels, each panel displaying information about the biomolecular sequence data of interest in a different way on a single screen or page. For example, a first panel could show a graphical representation of the entire biomolecular sequence, or the portion of the sequence of interest, with the locations of particular subsequences of interest indicated. A second panel could show a more detailed graphical representation of all or a selected portion of the sequence represented in the first window, allowing a user to focus on a particular subsequence of interest. This second panel view could depict additional information, such as annotations, relating to the particular subsequences of interest. A third panel could show information graphically representing the confid-

ence level or origination, for example, of the biomolecular sequence data represented in one or more of the other panels. Additional panels on the same or additional screens could show, for example, the actual nucleotide or amino acid sequence of or relating to a selected subsequence of interest represented in one or more of the other panels, or other information relating to the biomolecular sequence data.

In one preferred embodiment, a graphical viewer in accordance with the present invention provides a graphical representation of all or a selected portion of an organism's genome with its individual loci indicated. The viewer allows the user to focus on a particular region or locus of interest and have it also be graphically represented with additional information, such as annotations. A graphical depiction of sequence coverage for the sequence regions represented in the viewer may also be provided.

The viewer may also provide for the display of related loci from other portions of the organism's genome (i.e., paralogs), and allows for the retrieval of information about the loci, such as actual nucleotide sequences or detailed annotations, from an associated relational database system. In addition, a graphical viewer in accordance with the present invention may provide for the graphical representation and comparison of multiple portions of the genome of one or more organisms based on a locus of interest and its corresponding paralogs and homologs (related loci from another organism's genome).

A graphical viewer in accordance with a preferred embodiment of the present invention preferably provides graphical representations of the genomic data in a plurality of panels, each panel displaying information about the genomic data of interest in a different way. In a particularly preferred embodiment of the invention, the graphical viewer has three main panels on a single screen: a legend viewer, which shows the entire por

tion of the genome under consideration; a target viewer, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest; and a sequence depth viewer, which contains graphical information illustrating the depth of coverage over the length of the genome portion under consideration.

In one aspect, the present invention provides a method implemented in a computer system for presenting biomolecular sequence data. The method involves retrieving biomolecular sequence data from a database in response to a user query, and graphically depicting elements of the biomolecular sequence data in a user interface for the computer system. The graphical depiction may include a plurality of panels representing different aspects of the biomolecular sequence data in a single frame.

In a preferred embodiment, the biomolecular sequence data may include gene locus data and be graphically depicted in three panels, the first panel graphically depicting at least a portion of a contig and its associated loci, the second panel graphically depicting at least a portion of the contig depicted in the first panel and annotated loci associated with the portion, and the third panel graphically depicting information indicating the number of sequencing operations conducted to determine the sequence data depicted in the second panel. The third panel may graphically depict sequences used to assemble the portion of the contig depicted in the second panel, or depth of coverage information for the portion of the contig depicted in the second panel.

In another aspect, the invention provides another method implemented in a computer system for presenting biomolecular sequence data. The method involves retrieving biomolecular sequence data for a plurality of homologous loci from a database in response to a user query, and graphically depicting at least some of the homologous loci in a user interface for the computer system.

In yet another aspect, the invention provides a computer system. The computer system includes a database including biomolecular sequence data, and a user interface. The user interface is capable of receiving a query relating to the biomolecular sequence data, and graphically displaying the results of the query.

In still another aspect, the invention provides a computer-readable medium containing programmed instructions arranged to graphically display biomolecular sequence data. The computer-readable medium includes programmed instructions for retrieving biomolecular sequence data from a computer system database in response to a user query, and graphically depicting elements of the biomolecular sequence data in a user interface for the computer system.

These and other features and advantages of the present invention will be presented in more detail in the following specification of the invention and the accompanying figures which illustrate by way of example the principles of the invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Reference will now be made in detail to preferred embodiments of the invention. Examples of the preferred embodiments are illustrated in the accompanying drawings. While the invention will be described in conjunction with these preferred embodiments, it will be understood that it is not intended to limit the invention to one or more preferred embodiments. On the contrary, it is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well known process operations have not been described in detail in

order not to unnecessarily obscure the present invention.

Introduction

The present invention provides methods, media and systems for graphically displaying computer-based biomolecular sequence information. Generally, biomolecular sequence information may be graphically depicted in a variety of different forms in accordance with the present invention. The sequence information may be composed of nucleotide or amino acid sequence information or both. The graphical depictions may be in several different formats providing different information relating to the sequences, and may be displayed in one or more screens of a computer user interface.

A graphical viewer in accordance with the present invention preferably has as a plurality of panels, each panel displaying information about the biomolecular sequence data of interest in a different way on a single screen or page. For example, a first panel could show a graphical representation of the entire biomolecular sequence, or the portion of the sequence of interest, with the locations of particular subsequences of interest indicated. A second panel could show a more detailed graphical representation of all or a selected portion of the sequence represented in the first window, allowing a user to focus on a particular subsequence of interest. This second panel view could depict additional information, such as annotations, relating to the particular subsequences of interest. A third panel could show information graphically representing the confidence level or origination, for example, of the biomolecular sequence data as represented in one or more of the other panels. Additional panels on the same or additional screens could show, for example, the actual nucleotide or amino acid sequence of or relating to a selected subsequence of interest represented in one or more of the other panels, or other information relating to the biomolecular sequence data.

In one preferred embodiment, a graphical viewer in accordance with the present invention provides a graphical representation of all or a selected portion of an organism's genome with its individual loci indicated. The viewer allows the user to focus on a particular region or locus of interest and have it also be graphically represented with additional information, such as annotations. A graphical depiction of sequence coverage for the sequence regions represented in the viewer may also be provided.

The viewer also may provide for the display of related loci from other portions of the organism's genome (i.e., paralogs), and allows for the retrieval of information about the loci, such as actual nucleotide sequences or detailed annotations, from an associated relational database system. In addition, a graphical viewer in accordance with the present invention may provide for the graphical representation and comparison of multiple portions of the genome of one or more organisms based on a locus of interest and its corresponding paralogs and homologs (related loci from another organism's genome).

A graphical viewer in accordance with a preferred embodiment of the present invention preferably provides graphical representations of the genomic data in a plurality of panels, each panel displaying information about the genomic data of interest in a different way. In a particularly preferred embodiment of the invention, the graphical viewer has three main panels on a single screen: a legend viewer, which always shows the entire portion of the genome under consideration; a target viewer, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest; and a sequence depth viewer, which contains graphical information illustrating the depth of coverage over the length of the genome portion under consideration.

Of course, as noted above, a graphical viewer in accordance with the pre

sent invention may be used to display biomolecular sequence information other than the gene locus information described with reference to the preferred embodiments of the invention described herein. For example, a graphical viewer in accordance with the present invention may be used to display peptide or nucleotide sequence information, and can be used to display actual sequences resulting from comparisons of sequences from, for example, a BLAST or FASTA search.

The Graphical Viewer Environment

As noted above, a graphical viewer in accordance with the present invention is preferably used in connection with a biomolecular sequence relational database system, such as those developed by Lucyte Pharmaceuticals, Inc. of Palo Alto, CA, and described, for example, in United States Patent Nos. 5,970,500; 5,953,727; 5,966,712 and 6,023,659. Data to be displayed by a graphical viewer in accordance with the present invention is accessed from such a database system using techniques and commands well known to those of skill in the art. Figures 1A and 1B and the associated description provided below provide a context in which a graphical viewer in accordance with the present invention may operate.

Figure 1A depicts a network system 130 suitable for storing and retrieving information in relational databases, such as those suitable for supporting a graphical viewer in accordance with the present invention. Network 130 includes a network cable 134 to which a network server 136 and clients 138a and 138b (representative of possibly many more clients) are connected. Cable 134 is also connected to a firewall/gateway 140 which is in turn connected to the Internet 142.

Network 130 may be any one of a number of conventional network systems, including a local area network (LAN) or a wide area network (WAN), as is known in the art (e.g., using Ethernet, IBM Token Ring, or the like).

The network includes functionality for packaging client calls in a well

l-known format (e.g., URL) together with any parameter information into a format (of one or more packets) suitable for transmission across a cable or wire 134, for delivery to database server 136.

Server 136 includes the hardware necessary for running software to (1) access database data for processing user requests, and (2) provide an interface for serving information to client machines 138a and 138b. In a preferred embodiment, depicted in Figure 1A, the software running on the server machine supports the World Wide Web protocol for providing page data between a server and client. In this embodiment, a web server 156 having URL and HTTP functionality communicates with a client via the HTTP protocol.

Client/server environments, database servers, relational databases and networks are well documented in the technical, trade, and patent literature. For a discussion of database servers, relational databases and client/server environments generally, and SQL servers particularly, see, e.g., Nath, A., The Guide To SQL Server, 2nd ed., Addison-Wesley Publishing Co., 1995.

As shown, server 136 includes an operating system 150 (e.g., UNIX) on which runs a relational database management system 152, a World Wide Web application 154, and a World Wide Web server 156. The software on server 136 may assume numerous configurations. For example, it may be provided on a single machine or distributed over multiple machines.

World Wide Web application 154 includes the executable code necessary for generation of database language statements (e.g., Standard Query Language (SQL) statements). Generally, the executables will include embedded SQL statements. In addition, application 154 includes a configuration file 160 which contains pointers and addresses to the various software entities that comprise the server as well as the various external and internal databases which must be accessed to service user requests. Conf

figuration file 160 also directs requests for server resources to the appropriate hardware -- as may be necessary should the server be distributed over two or more separate computers.

Each of clients 138a and 138b includes a World Wide Web browser for providing a user interface to server 136, and including code necessary to generate HTML pages. Through the Web browser, clients 138a and 138b construct search requests for retrieving data from a sequence database 144 and/or a genomic database 146, for example. Thus, the user will typically point and click to user interface elements such as buttons, pull down menus, scroll bars, etc. conventionally employed in graphical user interfaces. The requests so formulated with the client's Web browser are transmitted to Web application 154 which formats them to produce a query that can be employed to extract the pertinent information from sequence database 144 or genomic database 146.

In the embodiment shown, the Web application accesses data in genomic database 146 by first constructing a query in a database language (e.g., Sybase or Oracle SQL). The database language query is then handed to relational database management system 152 which processes the query to extract the relevant information from database 146. In the case of a request to access sequence database 144, Web application 154 directly communicates the request to that database without employing the services of database management system 152.

The procedure by which user requests are serviced is further illustrated with reference to Figure 1B. In this embodiment, the World Wide Web server and/or executable Web application components of server 136 provide Hypertext Mark-up Language documents ("HTML pages") 164 to a client machine. At the client machine, the HTML document provides a user interface 166 which is employed by a user to formulate his or her requests for access to database 146. That request is converted by the Web application

a component of server 136 to a SQL query 168. That query is used by the database management system component of server 136 to access the relevant data in database 146 and provide that data to server 136 in an appropriate format. Server 136 then generates a new HTML document, possibly through the Web application 154, relaying the database information to the client as a view in user interface 166.

While the embodiment shown in Figure 1A employs a World Wide Web server and World Wide Web browser for a communication between server 136 and clients 138a and 138b, other communications protocols will also be suitable. For example, client calls may be packaged directly as SQL statements, without reliance on Web application 154 for a conversion to SQL. Clients may also query the database directly without using a client browser.

When network 130 employs a World Wide Web server and clients, it must support a TCP/IP protocol. Local networks such as this are sometimes referred to as "Intranets." An advantage of such Intranets is that they allows easy communication with public domain databases residing on the World Wide Web (e.g., the GenBank World Wide Web site). Thus, in a particular preferred embodiment of the present invention, clients 138a and 138b can directly access data (via Hypertext links for example) residing on Internet databases using a HTML interface provided by Web browsers and Web server 156.

Bear in mind that if the contents of the local databases are to remain private, a firewall 140 must preserve in confidence the contents of a sequence database 144 and a genomic database 146.

In a preferred embodiment, sequence database 144 is a flat file database with a single file for genomic sequences from different species. Other possible approaches may include partitioning the sequence data according to different species or whether or not sequences have been found to

be unique to the local database (i.e., sequences that did not have any hits in an external database such as GenBank).

Preferably, the information in genomic database 146 is stored in a relational format. Such a relational database supports a set of operations defined by relational algebra. It generally includes tables composed of columns and rows for the data contained in the database. Each table has a primary key, being any column or set of columns the values of which uniquely identify the rows in the table. The tables of a relational database may also include a foreign key, which is a column or set of columns the values of which match the primary key values of another table. A relational database is also generally subject to a set of operations (select, project, product, join and divide) which form the basis of the relational algebra governing relations within the database. As noted above, relational databases are well known and documented (see, e.g., Nath, A., The Guide To SQL Serve, referenced above).

A relational database may be implemented in different ways. In Oracle (trademark) databases, for example, the various tables are not physically separated, as there is one instance of work space with different ownership specified for different tables. In Sybase(trademark) databases, in contrast, the tables may be physically segregated into different "databases."

One specific configuration for network 130 for multiple users provides both the genomic and sequence databases on the same machine. If there is a high volume of sequence searching, it may be desirable to have a second processor of similar size and split the application across the two machines to improve response time.

A suitable dual processor server machine may be any of the following workstations: Sun - Ultra-Sparc 2(trademark) (Sun Microsystems, Inc. of Mountain View, CA), SGI - Challenge L(trademark) (Silicon Graphics, Inc.

of Mountain View, CA), and DEC - 2100A(trademark) (Digital Electronics Corporation of Maynard, MA). Multiprocessor systems (minimum of 4 processors to start) may include the following: Sun - Ultra Sparc Enterprise 4000(trademark), SGI Challenge XL(trademark), and DEC 8400(trademark). Preferably, the server machine is configured for network 130 and supports TCP/IP protocol.

Depending upon the workstation employed, the operating system may be, for example, one of the following: Sun Sun OS 5.5 (Solaris 2.5), SGI IRIX 5.3 (or later), or DEC Digital UNIX 3.2D (or later).

Databases used in conjunction with this invention may be downloaded via a 4 X 4 Gb+ FWSCSI 2, Fiber Link Raid Units 20Gb+, or 4 DAT Tape Drive. A CD ROM drive may also be acceptable.

The client machine may be, for example, a Macintosh(trademark) (Apple Computer Inc. of Cupertino, CA), a PC, or a Unix workstation. It should also be TCP/IP capable with a Netscape or Internet Explorer Web Browser.

The network may include a 10Base-T, 100Base-T or higher connection, be TCP/IP capable, and provide access to Internet for HTML hyperlinks to external databases.

Figure 1C illustrates the accessibility of graphical viewer features in accordance with a preferred embodiment of the present invention. A graphical viewer in accordance with the present invention is preferably provided together with a suite of functions made available to users through a collection of user interface screens (e.g., HTML or Java(registered trademark) pages) viewed in the user interface of a biomolecular relational database. Typically, the interface will have a main viewer page from which various lines of query can be followed. In a preferred embodiment, the main viewer page (and other graphical viewers) are Java(registered trademark)-based applets running on the network system. Given the func

tionalities described herein, one of ordinary skill in the art would be able to implement the graphical viewers of the present invention in Java (registered trademark) or other programming environments. The viewer page is typically accessed from another page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used.

For example, a user interface screen (e.g., HTML page) 170 displays textual information relating to a plurality biomolecular sequences. One or more sequences displayed in the page 170 may be selected, for example, using the pointer provided in the GUI, to access another page 180 which displays additional information about the selected sequences. This page 180 may include a button which when selected accesses a main graphical viewer page 190. The graphical viewer page (e.g., Java(registered trademark) page) 190 graphically depicts information about the selected sequences. The page also preferably includes buttons 192 which allow a user to modify the graphical display. The buttons 192 may also include buttons which a user may select to access additional graphical viewer pages 194, 196, which graphically or otherwise display additional information relating to the graphically displayed sequence information in page 190.

Gene Locus Implementation

The invention will now be described with reference to a particular preferred implementation of the invention to graphically depict gene locus information. The invention will be described with reference to a database optimized for microbial data, such as that described in United States Patent No. 5,970,500. However, application of the present invention is by no means so limited. For example, the invention covers graphical viewers used in connection with databases optimized for other sources of biomolecular sequence data, such as animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.) and plant sequences.

As noted above, a graphical viewer in accordance with the present invention is preferably provided together with a suite of functions made available to users through a collection of user interface screens viewed in the user interface of a biomolecular relational database. A main viewer page is typically accessed from another page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used, in this case a microbial genomic database. Figure 2 depicts one such other page from the microbial genomic database. The Contig Results page 200 displays a list of loci (identified by their LocusIDs) for genes localized to a particular "contig" (a group of assembled overlapping sequences), contig SAU1c0039, of the genomic sequence of a microbial organism, in this case *Staphylococcus aureus*.

By clicking on a particular LocusID in Contig Results page 200, a user accesses a Locus Information page, such as depicted in Figure 3. Clicking on the LocusID SAU100241 in page 200, returns the Locus Information page 300 which displays details about the locus SAU100241. The page also displays a Graphical Viewer button 302 which when selected launches a graphical viewer in accordance with the present invention.

Figure 4A depicts a main graphical viewer page 400 accessed by selecting the Graphical Viewer button 302 in Locus Information page 300. In this preferred embodiment, the graphical viewers are Java(registered trademark)-based applets that provide a graphical representation of a portion of a contig and its related loci. A graphical viewer in accordance with the present invention preferably includes a plurality of separate component viewers. Where more than one component viewer is featured it is preferably displayed in a single frame in order to enhance the effectiveness with which the graphically displayed data is conveyed to the user. A preferred embodiment includes three component viewers displayed in a sin

gle frame.

Thus, the graphical viewer 402 of page 400 has three viewer component panels on a single screen. The top panel 410 features a "legend viewer" 412, which shows the entire portion of the genome under consideration. The middle panel 430 features a "target viewer" 432, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest. The bottom panel 450 features a "sequence depth viewer" 452, which contains graphical information illustrating the depth of coverage over the length of the genome portion represented in the target viewer 422.

The graphical viewer page 400 also includes several buttons and windows along the top 403 of the page 400 for accessing and displaying additional information. A menu bar 404 is also provided for accessing pull-down menus listing various command and control functions. A scale 415, 435, 455 depicted at the bottom of each viewer panel 410. The use of these features will be described in further detail below.

The legend viewer 412 always shows the entire portion of the contig which was loaded by the viewer when the user selected a contig in the previous screens. In a preferred embodiment, the viewer will load a predetermined default number of base pairs of the contig sequence. If the contig is shorter than the default, the entire contig will be depicted and the default will be adjusted. For example, in this embodiment, the viewer loads 30,000 base pairs starting at the first locus in the list on the Contig Results screen 200 (identified by its Hit ID), g2462967. The number of base pairs shown and the position on the contig may be determined with reference to the scale 415 depicted at the bottom of the legend viewer panel 410. The default value may, of course, be changed to any desired number.

The legend viewer 412 graphically represents contig SAUlc0039 as a line 414 which starts at coordinate (base pair number) 4467 and extends up to

coordinate 34,467, as may be seen with reference to the scale 415. The contig depicted in the viewer is identified in a ContigID window 407. In addition, the starting coordinate for the portion of the contig depicted by the legend viewer 412 (namely, the starting coordinate of the selected locus g2462967: 4467) is noted in the Start Coord window 405. These windows 405, 407 may also be used to enter information in order to control the information depicted by the viewer, as described further below. A user may bring upstream or downstream portions of the contig into view in the legend viewer 412, and the other component viewers, by clicking on the directional buttons 406.

In addition to the contig, the legend viewer 412 shows the various loci residing on the portion of the contig. The manner in which these loci are depicted illustrates the power of a graphical viewer in accordance with the present invention in presenting information in a highly effective manner.

The loci are represented by arrows 416. Each loci is located beside the contig line 414 according to its position on the contig and the direction in which it is read. The arrowhead represents the direction in which a locus is read. Loci which are read in the forward (+) direction are depicted above the contig line 414. Loci which are read in the reverse (-) direction are depicted below the contig line 414. In addition, other graphical features may be used to convey information about the graphically depicted loci. For example, loci for which the sequences obtained are above an established confidence threshold may be depicted as broken arrows.

In this preferred embodiment, the loci are also represented in different colors based on their protein's function. Proteins are grouped into various functional categories, with each category being assigned a color.

For example, in this preferred embodiment, the proteins corresponding

to loci are grouped according to the following categories/colors: Motility/Light blue; Virulence/Red; Transport/Light Green; Regulation/Magenta; Macromolecule metabolism/Yellow; Small molecule metabolism/Dark blue; Structure/Dark Green; and Unclassified/Black. Of course other categories and colors may also be used. These arrow and color representation features for loci are used in both the legend viewer and the target viewer, discussed below.

The target viewer 432 initially displays the same scope as the legend viewer 412. The scope of the target viewer may be modified, however, by clicking on the Zoom buttons 409. The Zoom In button provides a closer view of a portion of the contig shown in the legend viewer 412. The closer view is depicted in the target viewer 432, with the scale 435 adjusting to reflect the amount of the zoom. The Zoom Out button provides a broader view of the contig, up to the maximum of the default base pair number selected for the legend viewer (minimum magnification). The Zoom None button automatically returns to the minimum magnification.

Another way provided by a graphical viewer in accordance with the present invention to focus on a portion of interest of a contig 414 depicted in the legend viewer 412 is to provide an outline, such as a colored (e.g., red) box, around the portion of the contig 414 which is shown in the target viewer 432. In this preferred embodiment, when a red box surrounds the entire legend viewer panel, the target viewer also displays the entire 30,000 base pairs. This is the situation illustrated in Figure 4A.

When the Zoom buttons 409 are used, as described above, the red box is adjusted accordingly.

An area on the contig may also be zoomed into by direct user adjustment of the red box (known as "rubber banding"). The scope of the red box may be changed by clicking at a location in any of the viewer panels and dragging the cursor with a mouse to another location. The red box will t

then encompass the region between those two points, and only this region will be visible in the target and sequence depth viewers. Figure 4B depicts an updated page showing the viewer 402 after a user has zoomed in on the portion 434 of the contig 414 depicted in the legend viewer 412 between about the coordinates 14,200 and 18,200. The scale 435 at the bottom of the target viewer 432 has been adjusted to reflect the new scope of the zoomed target view.

Another feature of the target viewer is the loci are annotated. As may be seen in Figures 4A and 4B, annotations 436 are provided for loci arrows which are long enough to accommodate the annotation information. If a loci of interest is too short to be display its annotation, a user may zoom in further on the locus until it is long enough to allow the annotation to be displayed in the graphical representation.

Individual loci in the target viewer 432 may be selected for further analysis by clicking on the graphically depicted locus. A selected locus is highlighted in some manner, for example, by displaying a colored (e.g., red) box around its representation. Details about this locus may be viewed by double-clicking on the locus representation. Double-clicking opens a Selected Object Details window, such as depicted in Figure 5A. The Selected Object Details window 500 includes information about the locus, including its LocusID, gene (functional) category, base pair range, the sequence's homologous matches (preferably the number of homologous matches returned is limited to a preset number; for example, the top five matches are returned here) against other sequence databases, for example, the genpept database, and other information useful to researchers and relating to other features of the database system with which the graphical viewer is used. Many of the fields of information provided in the window 500 may be hyperlinks to other HTML pages or other screens.

The Selected Object Details window 500 includes an Alignment button 502.

Clicking on this button accesses an alignment viewer which provides a graphical representation of the locus sequence and its homologous matches. An example of an alignment viewer 510 in accordance with a preferred embodiment of the present invention is shown in Figure 5B. The alignment viewer 510 has three panels. The top two panels 512 and 514 provide a graphical representation of the locus identified in Figure 5A (SAU101156). The third panel 516 provides graphical representations of the five homologs noted in Figure 5A. The alignment viewer page also includes a number of buttons 518 which may be used to control the graphical representations. In particular, the page has Zoom buttons 520 which may be used to zoom into the sequence level of loci depicted in the lower two panels 514 and 516 (while the upper panel 512 maintains the depiction of the entire locus). Figure 5C illustrates this Zoom feature where the upper panel 512 has a colored box 522 around the portion of the locus depicted with its homologs at the sequence level in the two lower panels 514 and 516. In this embodiment, the amino acid sequences are shown. In other embodiments, the corresponding nucleotide sequences may also be shown.

An additional feature of the graphical viewer page 400 that becomes useful when the scope of the view in the target viewer 432 is focused in on a portion of the contig sequence shown in the legend viewer 412 is a scrollbar 470 at the bottom of the page. The scrollbar 470 allows a user to move along the portion 434 of the contig 414 to bring upstream or downstream portions of the contig 414 into view in the target viewer 432. The third panel 450 of the graphical viewer 402 in this embodiment of the present invention is the sequence depth viewer 452. The sequence depth viewer 452 provides a graph illustrating the depth of coverage, that is, the number of times that a given portion of the contig has been sequenced, over the length of the contig. The sequence depth viewer 452 disp

lays its graph for the contig or portion of the contig displayed in the target viewer 432. Thus, in Figure 4A, were the target viewer 432 and legend viewer 412 have the same scope, the sequence depth viewer 452 displays a graph showing the depth of coverage over the 30,000 base pairs of the contig 414 from coordinates 4467 to 34,467, as indicated by the scale 455 at the bottom of the sequence depth viewer panel 450. In Figure 4B, however, the sequence depth viewer 452 displays a graph showing the depth of coverage over the approximately 4000 base pairs of the portion 434 of the contig zoomed in on in the target viewer from about coordinates 14,200 to 18,200, as indicated by the adjusted scale 455. The sequence depth viewer also includes a second scale 456 on the y-axis indicating the number of sequencing passes represented by the graph.

The manner in which this depth of coverage information is depicted provides a further illustration of the power of a graphical viewer in accordance with the present invention in presenting information in a highly effective manner. A user of the graphical viewer is able to very quickly, at a glance, assimilate useful information relating to the confidence to be attributed to the sequence information depicted in the other panels of the viewer. In this preferred embodiment of the present invention, the sequence depth viewer 452 depicts coverage as a sequence distribution graph 454. A particular advantage of this way of depicting of the depth of coverage information is that it is particularly effective for clearly providing this information in a graphical format which makes a clear visual impression and renders the data easily quantifiable, with reference to the y-axis scale 456. The coverage data for various regions is also easily compared in this format.

In other embodiments of the invention, a sequence depth viewer may graphically depict depth of coverage information in other ways. For example, the actual sequences from which the contig was assembled may be depicted

d. This way of depicting the sequence coverage information may provide useful information for some users who are concerned with the data acquisition process, for example, used in the contig's formation.

As noted above, the graphical viewer page 400 includes several buttons and windows along the top 403 of the page 400 for accessing and displaying additional information. Several of these have already been discussed, including the Start Coord 405 and ContigID 407 windows. Figures 6 and 7 illustrate additional features of a this embodiment of a graphical viewer in accordance with the present invention.

In addition to displaying the start coordinate for the contig sequence displayed in the legend viewer 612, the Start Coord window 605 may receive an entry from a user of a different starting coordinate. The entry of a different start coordinate will bring a different portion of a contig's sequence into view in the legend viewer. For example, Figure 6 shows a graphical viewer page 600 with the same settings as page 400, except that 0 has been entered in the Start Coord window 605. As a result, the contig sequence 602 and associated loci 604 shown in the legend viewer 612 is shifted 4467 base pairs upstream to the beginning of contig SAUlc0039. The 4467-most downstream base pairs in the depiction of the contig 414 in Figure 4 are no longer visible in the viewer of page 600. The corresponding views are also depicted by the target viewer 632 and the sequence depth viewer 652.

Also, in addition to identifying the contig depicted in the viewer 402, the ContigID window 407 may receive an entry from a user of a different ContigID. The entry of a different ContigID will cause the default number of base pairs (preferably starting from the coordinate 0) of the contig sequence associated with the new ContigID to be loaded from the database associated with the viewer and displayed. For example, Figure 7 shows a graphical viewer page 700 with the ContigID SAUlc0016 entered in t

he ContigID window 707. As a result, the contig sequence 702 and associated loci 704 shown in the legend viewer 712 are that for contig SAUlc0016. The corresponding views are also depicted by the target viewer 732 and the sequence depth viewer 752.

As also noted above, the graphical viewer 400 includes a menu bar 404 for accessing pull-down menus listing various command and control functions. The File pull-down menu lists standard commands found in applications software packages such as save and print, etc. The Edit pull-down menu provides a list of categories for editing the parameters of the graphical viewers, including the default contig sequence length display number and the colors used to represent various features in the viewers.

Of particular interest is the View pull-down menu which, together with allowing the user to select which features should be included in the various viewer displays, also includes a View Sequences on Coords 804 option. A graphical viewer page 800 is shown in Figure 8A with the View pull-down menu 802 selected. Selection of the View Sequences on Coords 804 option from the menu 802 accesses a page 810 listing the sequences used to assemble the contig depicted in the graphical viewer 402, together with the coordinates of each sequence which indicate its coverage. Selecting a sequence from the list, such as the second one in the list, 806503054F1 (5201,5690) 812, and clicking the Sequence Database button 814 accesses a database of the raw sequences used to assemble contigs in the database system associated with the graphical viewer and returns a Sequence Retrieval Results page 820, depicted in Figure 8C. The Sequence Retrieval Results page 820 depicts the actual nucleotide sequence 822 of the sequence 812 selected in Figure 8B.

A generalized process by which a graphical viewer system in accordance with a preferred embodiment of the present invention returns graphical representations of gene locus information to a user is depicted in Figure

9. This process flow shows only some of the main features of a preferred embodiment of the present invention in order to illustrate in process flow form some of the options for graphically displaying sequence data in accordance with an embodiment of the present invention. It is not intended to provide a comprehensive depiction of the present invention.

The process 900 begins at 901 and at a step 902 data for a selected locus and its associated contig are loaded into the graphical viewer. As noted above, the locus may be selected from a list in a HTML page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used, in this case a microbial genomic database. At a step 904, a graphical display of the selected locus on its contig is provided. Preferably, the graphical display has a plurality of components for representing different aspects of the biomolecular sequence data associated with the selected locus.

In a particularly preferred embodiment depicted in Figures 4A and 4B and described above, the graphical representation is a viewer having three components: a legend viewer, a target viewer, and a sequence depth viewer.

If no further entries or zoom adjustments are made, the process may end at step 940 following the graphical display at step 904. However, a user may want to use the graphical viewer to extract and display additional information relating to the selected locus or other loci, and the viewer provides additional functionalities for this purpose.

The graphical representation of the data displayed by the graphical viewer may be modified in a variety of ways. Also, additional information may be accessed by selecting various objects (namely, loci) in a viewer.

For example, a user may enter a new contigID in a field provided in a graphical viewer page, such as window 407 in Figure 4A. If so, decision step 906 is answered in the affirmative and the new contig and its loci

are graphically depicted in a viewer at a step 914. A user may also enter a new start coordinate, such as in Start Coord window 405 in Figure 4A. If so, decision step 908 is answered in the affirmative and the graphical display is adjusted to show the contig in the new coordinate range at a step 916. In addition, as described above, a user may choose to focus in on a particular portion of a graphically depicted contig. If so, decision step 910 is answered in the affirmative and the graphical display in the target viewer, in this embodiment, is adjusted to show the contig in the zoomed view at a step 918. If any of these decision steps are answered in the negative the graphical viewer display remains unchanged.

After any of these actions, or in the alternative, a user may select an object to obtain further information. In a preferred embodiment, the loci depicted in a target viewer component of the graphical viewer may be selected by clicking on its representation. If so, decision step 920 is answered in the affirmative and the depiction of the locus in the target viewer may be highlighted with a colored box. If a user wishes to obtain detailed information about the selected loci, the user may do so by double clicking on the depiction of that locus. If so, decision step 921 is answered in the affirmative and a Java(registered trademark) page showing detailed information about the selected locus is shown at a step 922.

Another feature of this aspect of a preferred embodiment of the present invention is a graphical alignment viewer, as described above. A user may elect to display a graphical viewer which shows the alignment of the amino acid sequence of the loci of interest against some 'homologous sequences. If so, decision step 924 is answered in the affirmative and the alignment is graphically displayed in a graphical viewer at a step 926. A user may also be provided with the option of displaying a multiple org

anism viewer to view graphical representations of homologous and paralogous loci of the locus of interest. For example, if decision step 920 is answered in the affirmative, a multiple organism viewer may be accessed at a step 930 when a decision step 928 is answered in the affirmative.

Further details of the operation of a multiple organism viewer in accordance with a preferred embodiment of the present invention are described below with reference to Figures 10A-10E and 11.

Of course the selected object details and multiple organism selection decisions are independent of each other and could just as easily have been presented in other ways in Figure 9. Further, it should be noted that the system allows the user to exit from the graphical viewer mode at any time. This option is not depicted in Figure 9.

A further option available for accessing further information from a graphical viewer in accordance with the present invention is the display of actual nucleotide or amino acid sequences for a selected sequence associated with the locus of interest and its contig. In a preferred embodiment, a user may choose this option by clicking on a button in a graphical viewer page such as depicted in Figure 4A. If so, decision step 912 is answered in the affirmative and a list of sequences (sequence identifiers) and coordinates for the sequences from which the contig displayed in the viewer was assembled is displayed at a step 932. A user may then select a sequence from the list. If so, decision step 934 is answered in the affirmative and the actual nucleotide sequence (in this case) of the selected sequence is displayed. The process ends at 940.

As with other data displayed in graphical viewers in accordance with the present invention, the data used in this aspect of the invention is obtained from an associated biomolecular sequence database and system. The organization and operation of such systems may vary. Examples are provided in the Incyte Pharmaceuticals patents previously referred to herein.

n. Given the description of the functionality and displays herein, one of skill in the art would be able to implement the graphical viewer of the present invention in any such system.

Multiple Organism Viewer

As noted above, a graphical viewer in accordance with the present invention may also provide for the graphical representation and comparison of multiple portions of the genome of one or more organisms based on a locus of interest and its corresponding paralogs (related loci from other portions of an organism's genome) and homologs (related loci from another organism's genome). A preferred embodiment of such a multiple organism viewer is described with reference to Figures 10A-10D, below.

Figure 10A depicts a main graphical viewer page 1000, like that shown in Figures 4A and 4B. In Figure 10A, a box ("rubber band") 1002 has been placed around a region 1004 of the portion of the contig 1006 displayed by the legend viewer 1008 component of the graphical viewer 1010. This region 1004 of the contig 1006 is displayed by the target viewer 1020 component of the graphical viewer 1010, and its coverage is depicted by the sequence depth viewer 1030 component. In the target viewer 1020, a box 1022 around locus SAU100242 indicates that that locus has been selected. As noted previously, the main viewer page 1000 includes a Multiple Organisms button 1001.

Clicking on the Multiple Organisms button 1001 when a locus has been selected in the target viewer retrieves from the database associated with the viewer and displays a list of all libraries containing homologs and paralogs of the selected locus. Figure 10B depicts a window 1040 showing a list of libraries retrieved for the locus SAU100242 selected in the previous page shown in Figure 10A. To access a list of individual homologs and paralogs, a user may select one or more libraries in the list displayed in this window 1040. Clicking on the Multiple Organisms button 1

042 retrieves the individual homologs and/or paralogs and displays them.

Figure 10C depicts an example of a window 1045 showing a list of homologs and paralogs for the locus SAU100242 from the libraries selected in screen 1040 shown in Figure 10B. A hit description for each locus is also provided.

A user may then choose to produce a graphical display of the originally selected locus (e.g. SAU100242) and the selected homologous and paralogous loci displayed in the list of Figure 10C. By clicking on the Multiple Organisms button 1046 in window 1045, the locus of interest and its homologs and paralogs are loaded into a multiple organism viewer in accordance with a preferred embodiment of the present invention, and the locus of interest and the selected homologs and paralogs are displayed. Figure 10D depicts an example of such a multiple organism viewer page 1050.

The multiple organism viewer page 1050 provides a single panel multiple organism viewer 1052 graphically depicting the selected locus of interest (SAU100242) on its contig (SAU1c0039) and the selected homologous and paralogous loci on their respective contigs. Figure 10D, shows a viewer 1052 graphically displaying five (5) contigs in a single page: SAU1c0039 1061, PRT1c0129 1062, SAU2c0391 1063, SEP1c0220 1064, and SHA1c0122 1065. Contig 1061 is shown together with its loci, including the selected locus SAU100242 1071 depicted in bold and italicized in order to more clearly identify it. In the embodiment depicted in Figure 10D, the loci are annotated with a hit description rather than a Locus ID. Each of the other contigs is also depicted with its loci alongside, and with the loci homologous to SAU100242 (respectively, loci 1072, 1073, 1074 and 1075) shown in bold italics.

The multiple organism viewer 1052 illustrates another example of the power of a graphical viewer in accordance with the present invention to convey biomolecular sequence information in an effective way. As noted above,

ve, the selected locus and its homologous and paralogous loci may be shown bold and italicized, or in other type, such as a particular color, in order to distinguish them as the loci of interest. As may also be seen in Figure 10D, the loci of interest for the graphically displayed contigs are aligned in the page 1052 so that a visual comparison of adjacent loci on the various contigs is easily achieved. This visual representation may be further enhanced through use of the complement feature described below with reference to Figure 10E.

Further features of such a graphical viewer in accordance with this embodiment of the invention may be accessed by clicking on pull-down menu selections 1053 provided in the multiple organism viewer page 1050. The menu selections include File, View and Help selections that provide features such as described above with reference to Figure 4A. The Show selection 1054 accesses a list of all of the loci listed in the window illustrated in Figure 10C and loaded into the multiple organism viewer. By selecting a locus from the Show pull-down menu, a user may determine that the locus along with the contig on which it resides will be displayed or hidden. clicking on the loci, a user may determine that a locus will be displayed or hidden. The Show menu may also provide for the same determination to be made with respect to the contigs.

The Complement menu selection 1055 allows a user to manipulate the graphical representations of the contigs and loci in order to facilitate the extraction of salient information from the data. In particular, the complement menu selection 1055 allows the user to perform a reverse complement on any of the contigs displayed in the multiple organism viewer 1052. In this way, the homologous and paralogous loci displayed in the viewer 1052 may be shown in the same reading direction so that a user may more easily locate patterns of related loci adjacent to the loci of interest. Figure 10E depicts a multiple organism viewer page 1080 in which

the loci of interest depicted in page 1050 are shown with the same reading direction by use of the complement feature to show the reverse complement of contigs 1062, 1063 and 1065.

Shortcuts for the complement feature, as well as other features described herein, may be made available to a user according to methods well known to those of skill in the art. For example, the complement of a locus (contig) may be shown by holding down the shift key on a keyboard used to interface with the computer system on which the graphical viewer is operating while clicking on the contig.

Figure 11 depicts a flow chart for a generalized process of the operation in a multiple organism viewer in accordance with a preferred embodiment of the present invention. The process 1100 starts at 1101, and at a step 1102 the multiple organism viewer system receives a selection of a locus of interest, for example by clicking on a locus in the target viewer of Figure 10A. At a step 1104, a list of libraries containing loci homologous or paralogous to the selected locus of interest is displayed in a window. This display may be initiated by a user clicking on a button, such as the Multiple Organisms button in Figure 10A. Next, the system receives a selection of one or more libraries from the list at a step 1106, and at a step 1108 a list of loci from the selected libraries which are homologous or paralogous to the selected locus of interest is displayed in a window. At a step 1110, the system receives a selection of loci homologous or paralogous to the selected locus of interest from the list to be displayed. Then, at a step 1112, the selected loci and their respective contigs are graphically displayed in a multiple organism graphical viewer. In a preferred embodiment, the viewer shows all of the contigs and loci in a single panel in order to facilitate comparison of the graphically depicted data. The process ends at 1114.

Implementation

It is important to note that the present invention may be implemented as a system or a method, and may be embodied on a variety of computer-readable media that include program instructions, etc. for performing various operations described herein. As noted above, the system implementation is preferably in association with a biomolecular sequence relational database system. The method is a computer-implemented method, generally involving the operation of such a system. The media may be any computer-readable media. Examples of computer-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). The invention may also be embodied in a carrier wave travelling over an appropriate medium such as a microwave, optical lines, electric lines, etc.

Conclusion

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. It should be noted that there are many alternative ways of implementing methods, media and systems of the present invention. As noted previously, the scope of the invention is not limited to use with a microbial genomic database system such as that in connection with which the invention is primarily described above. Given the description provided herein, one of skill in the art would understand how to use the present invention in connection with a variety of computer-based biomolecular sequence database systems. For example, a graphical viewer in accordance with the present invention may be used in connection with database systems employed to store and analyze other types and forms of nucleic acid

sequences or expressed nucleic acid or amino acid sequences. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

4. BRIEF DESCRIPTION OF DRAWINGS

Figure 1A is a block diagram of a client-server Intranet for providing database services in accordance with one embodiment of the present invention.

Figure 1B is a schematic representation of the various software documents and entities employed by the Figure 1A client-server Intranet to provide biological information in response to user queries.

Figure 1C is a block diagram illustrating the accessibility of graphical viewer features in accordance with a preferred embodiment of the present invention in connection with a biomolecular sequence database.

Figure 2 is a screen shot (HTML page) depicting a Contig Results page for a graphical user interface of a genomic sequences database suitable for selecting a locus to be viewed with a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 3 is a screen shot depicting a Locus Information page for a graphical user interface of a genomic sequences database suitable for accessing a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 4A is a screen shot depicting a main page of a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 4B is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the zoom feature in accordance

ce with one embodiment of the present invention.

Figure 5A is a Selected Object Details window in accordance with one embodiment of the present invention.

Figures 5B and 5C are screen shots depicting an alignment viewer in accordance with one embodiment of the present invention.

Figure 6 is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the new starting coordinate feature in accordance with one embodiment of the present invention.

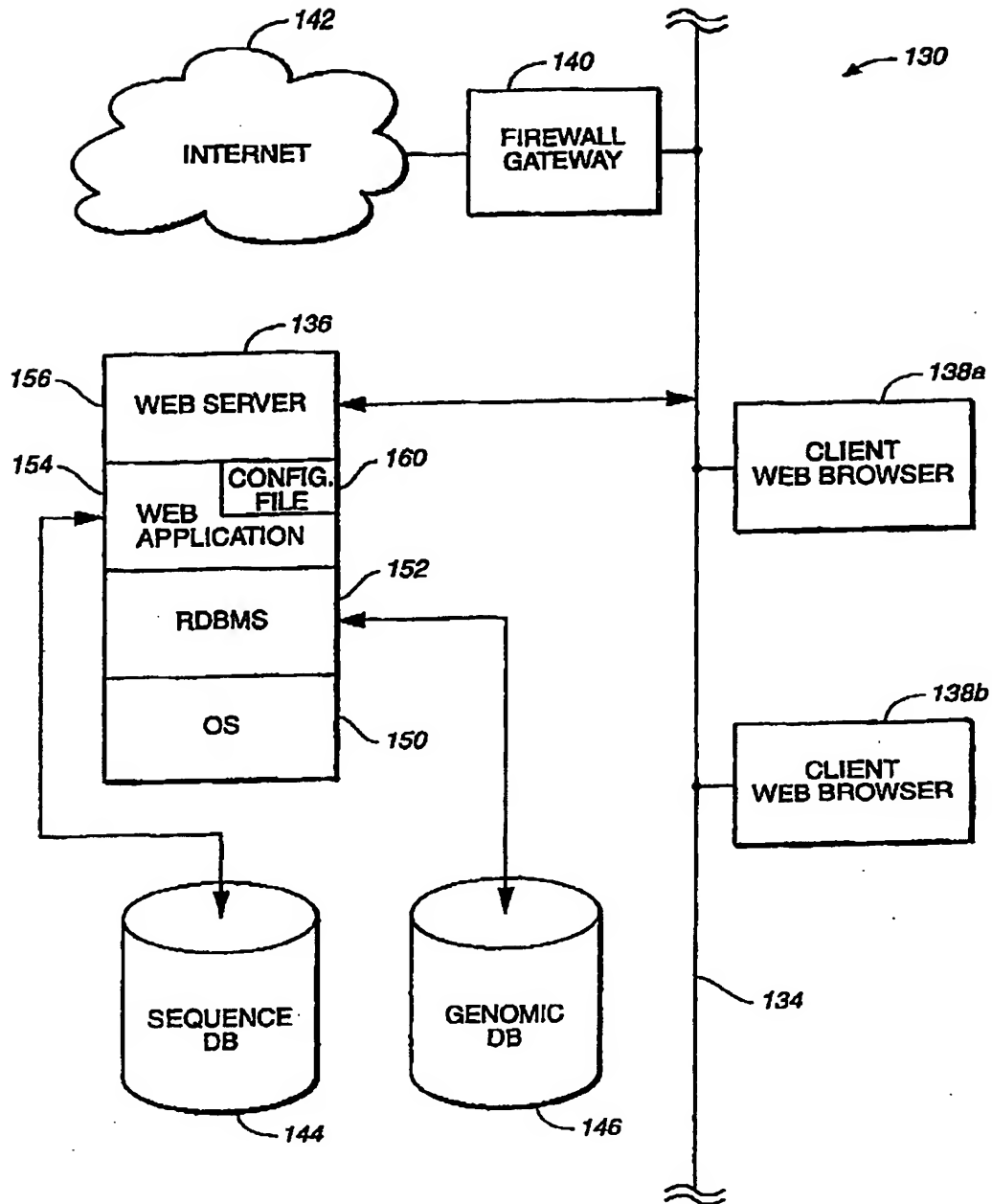
Figure 7 is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the new ContigID feature in accordance with one embodiment of the present invention.

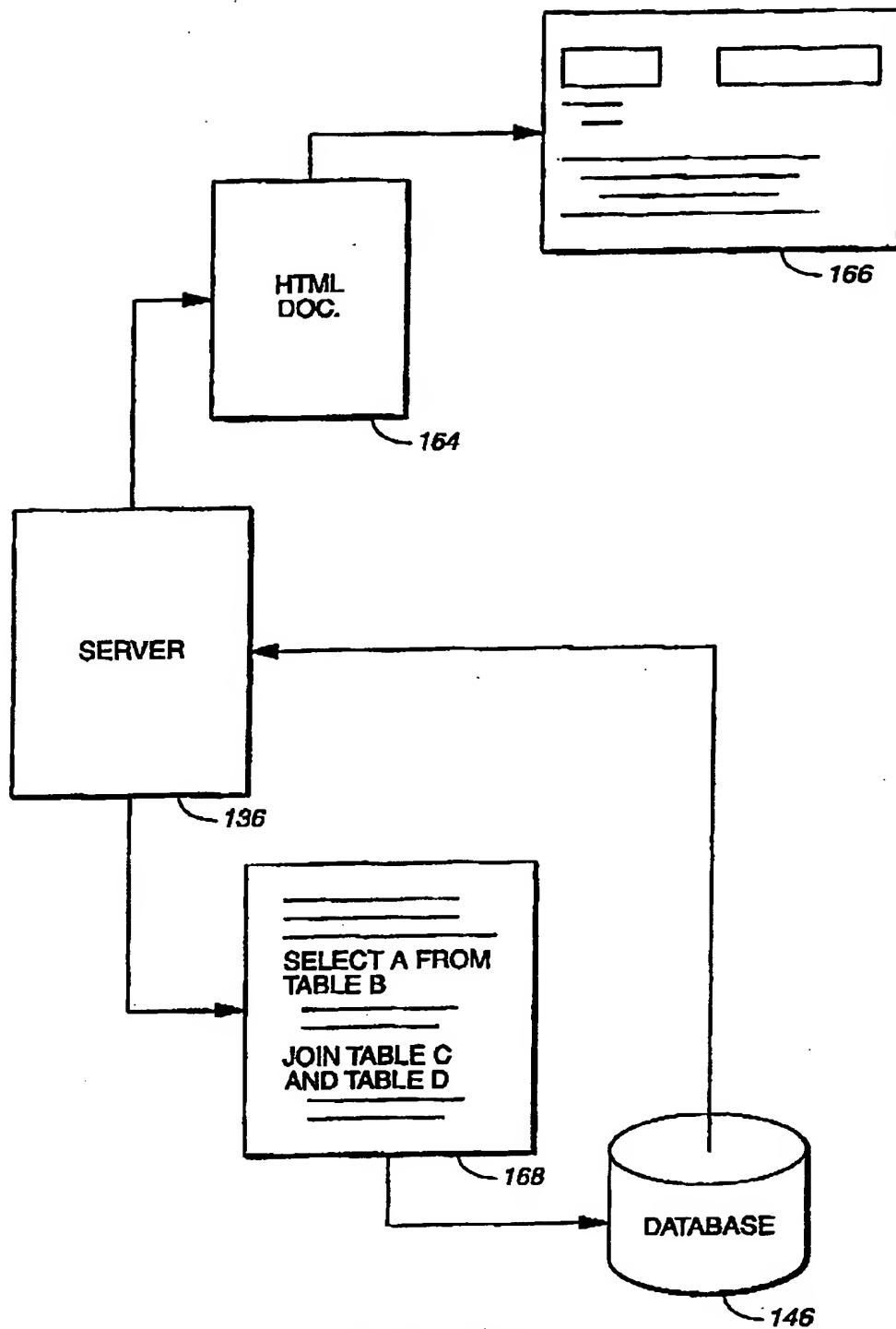
Figures 8A-8C are screen shots depicting pages of a biomolecular sequence graphical viewer illustrating a feature which displays an actual biomolecular sequence in accordance with one embodiment of the present invention.

Figure 9 is a flow chart depicting a process flow by which gene locus information may be viewed with a biomolecular sequence graphical viewer in accordance with a preferred embodiment of the present invention.

Figures 10A-10E are screen shots depicting the operation of a multiple organism biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 11 is a flow chart depicting a process flow by which multiple organism gene locus information may be viewed with a biomolecular sequence graphical viewer in accordance with a preferred embodiment of the present invention.

**FIG. 1A**

**FIG. 1B**

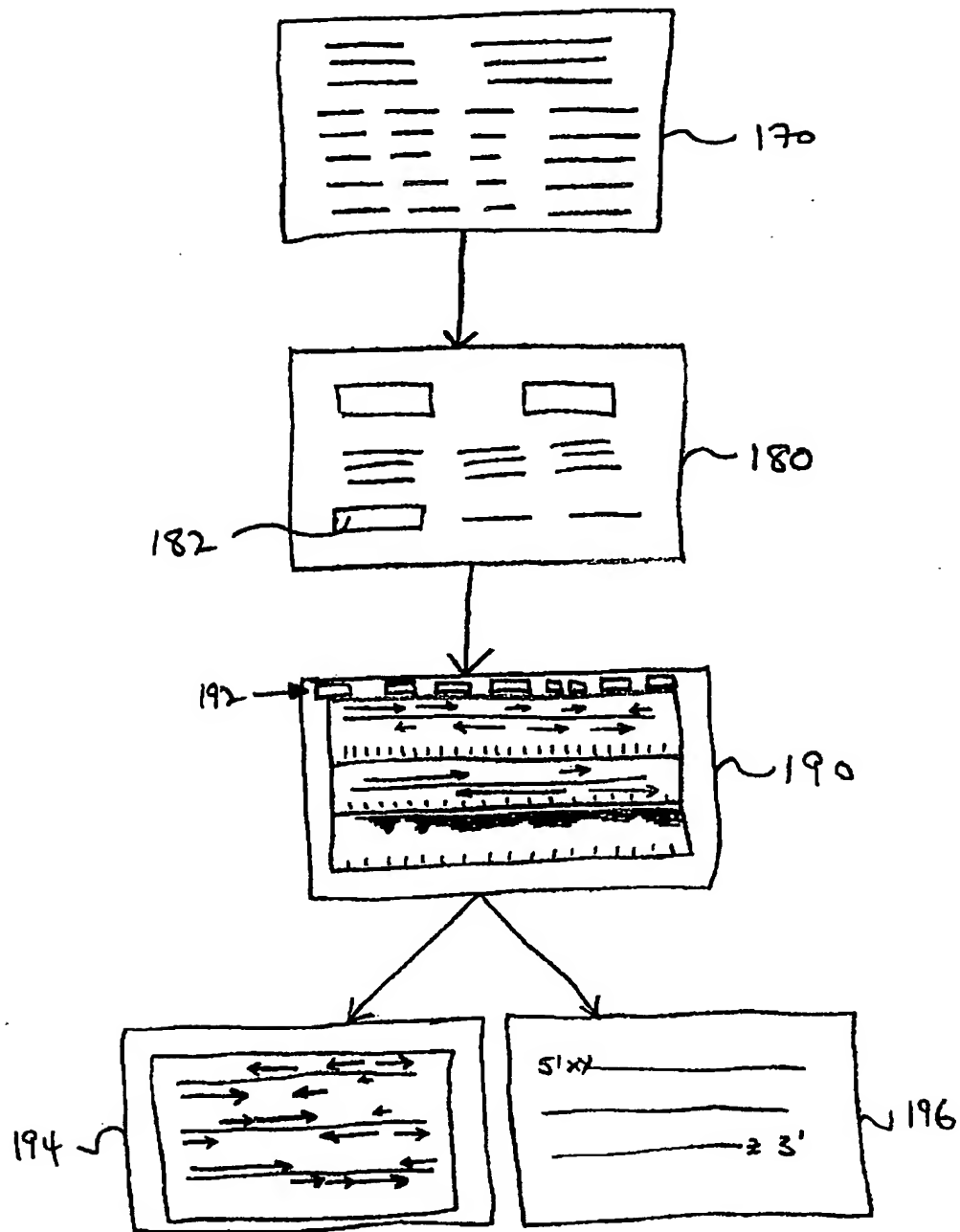


FIG. 1C

PathoSeq™ Genomics for Life™

PathoSeq™

PathoSeq™ Genomics for Life™

Contig Results

Plant Name	Org Info	Genome	Contig Number	Contig	Locs Info	Protein	Empirical	Sequences	Help
------------	----------	--------	---------------	--------	-----------	---------	-----------	-----------	------

Certificate of PathoSeq™ Genomics, Inc. PathoSeq Version 4.1 AP 9/99

Library: 5102001 Staphylococcus aureus Contigs: 48 Contig 002a: 134

Contig: 5A10C0039 Length: 138277 bp Segs: 1846 Base Pairs Selected: 1 to 138277

LocusID	Hit ID	Hit Description	Hit Organism	E-Value	Segs Pairs	Libs (48)	Segs
---------	--------	-----------------	--------------	---------	------------	-----------	------

5A10102730	02633184	Y280	Bacillus subtilis	2.4e-51	22-2169(-)	0	26
5A10102730	02633184	LOR	Bacillus subtilis	2.4e-51	2170-2722(+)	0	21
5A10102730	02633184	similar to hypothetical proteins	Bacillus subtilis	4.8e-62	2723-3359(+)	24	30
5A10102730	02633184	ribonuclease X	Bacillus subtilis	3.2e-39	3359-4332(+)	34	9
5A10102730	02633184	putative succinyl-CoA synthetase beta ch	Bacillus subtilis	6.8e-123	4467-5621(+)	35	11
5A10102730	02633184	succinyl-CoA synthetase (alpha subunit)	Bacillus subtilis	2.9e-97	5655-6542(+)	21	20
5A10102730	02633184	LysN	Staphylococcus	4.8e-21	6849-4998(+)	2	7
5A10102730	02633184	LysN	Staphylococcus	4.8e-21	7029-8270(+)	13	18
5A10102730	02633184	ORF4	Staphylococcus	2.5e-109	8314-9214(+)	36	11
5A10102730	02633184	DNA topoisomerase I	Bacillus subtilis	0	9497-11562(+)	40	21
5A10102730	02633184	Cid protein	Bacillus subtilis	0	11728-12026(+)	40	20
5A10102730	02633184	integrase/recombinase	Bacillus subtilis	7.1e-51	12440-14333(+)	30	27
5A10102730	02633184	hate-type subunit of the 20S proteasome	Bacillus subtilis	9.7e-32	14348-14981(+)	21	17
5A10102730	02633184	ATP-dependent Clp protease-like	Bacillus subtilis	7.1e-95	14971-16350(+)	40	16
5A10102730	02633184	transcriptional regulator	Bacillus subtilis	1.3e-63	16378-17139(+)	9	8
5A10102730	02633184	ribosomal protein S1	Bacillus subtilis	2.6e-73	17493-18109(+)	47	17

PathoSeq™ Genomics, Inc. 10000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

Fig. 2

200

PathoSeqTM
Genomics for life

Locus Information

Search By LocusID: SAU100241 Display: G Locus Details

C Locus around the selected Locus

Annotations Gene Coding Sequence Coding Initiating Protein Compound Compound Hit

Graphical View: 302

Locus ID: SAU100241 Type: Knowledge Gene Category: Small molecule metabolism
 Contig ID: SAU100019 Position: 2 of 150 Sequ: 13
 Accession: 385 Nucleotides: 1155 Lib: 33
 Homology: 16 Paralogs: 1

Top hits: 7A37A of ORF against papspept110

Hit ID	Hit Description	Hit Organism	E Value	ORF Coverage
0245297	putative succinyl-CoA synthetase beta ch	Bacillus subtilis	6.8e-113	100%
0131981	succinyl-CoA synthetase (beta subunit)	Bacillus subtilis	6.8e-113	100%

PathoSeqTM Document Type: 22/2/2001

300

FIG.3

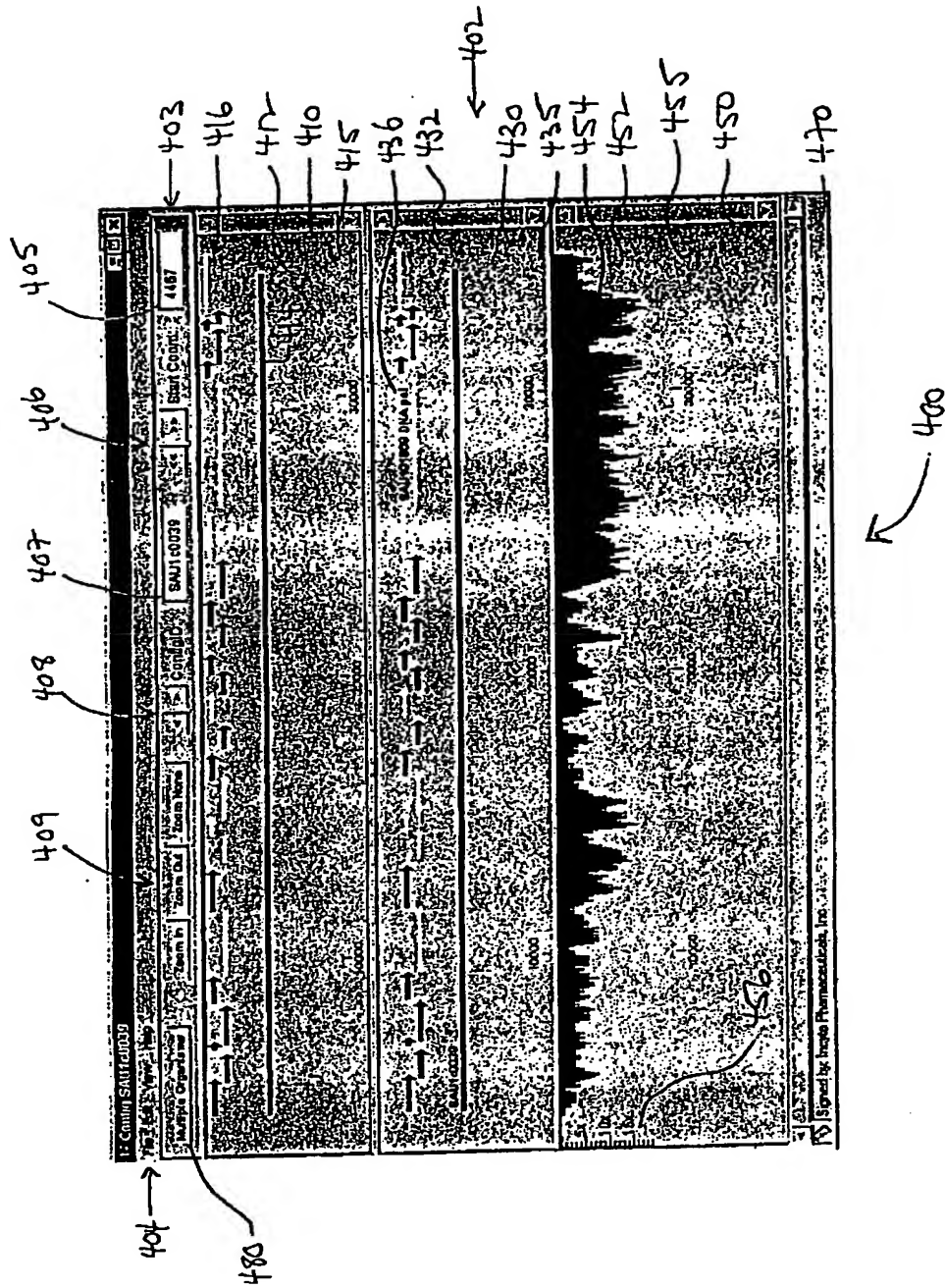


FIG. 4A

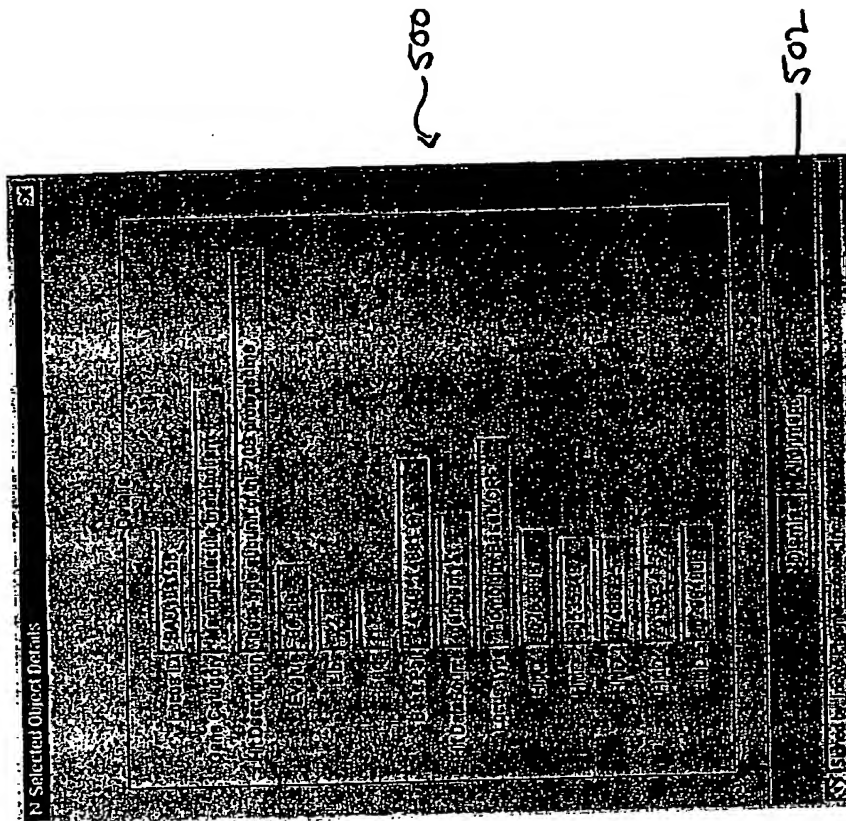


FIG. 5A

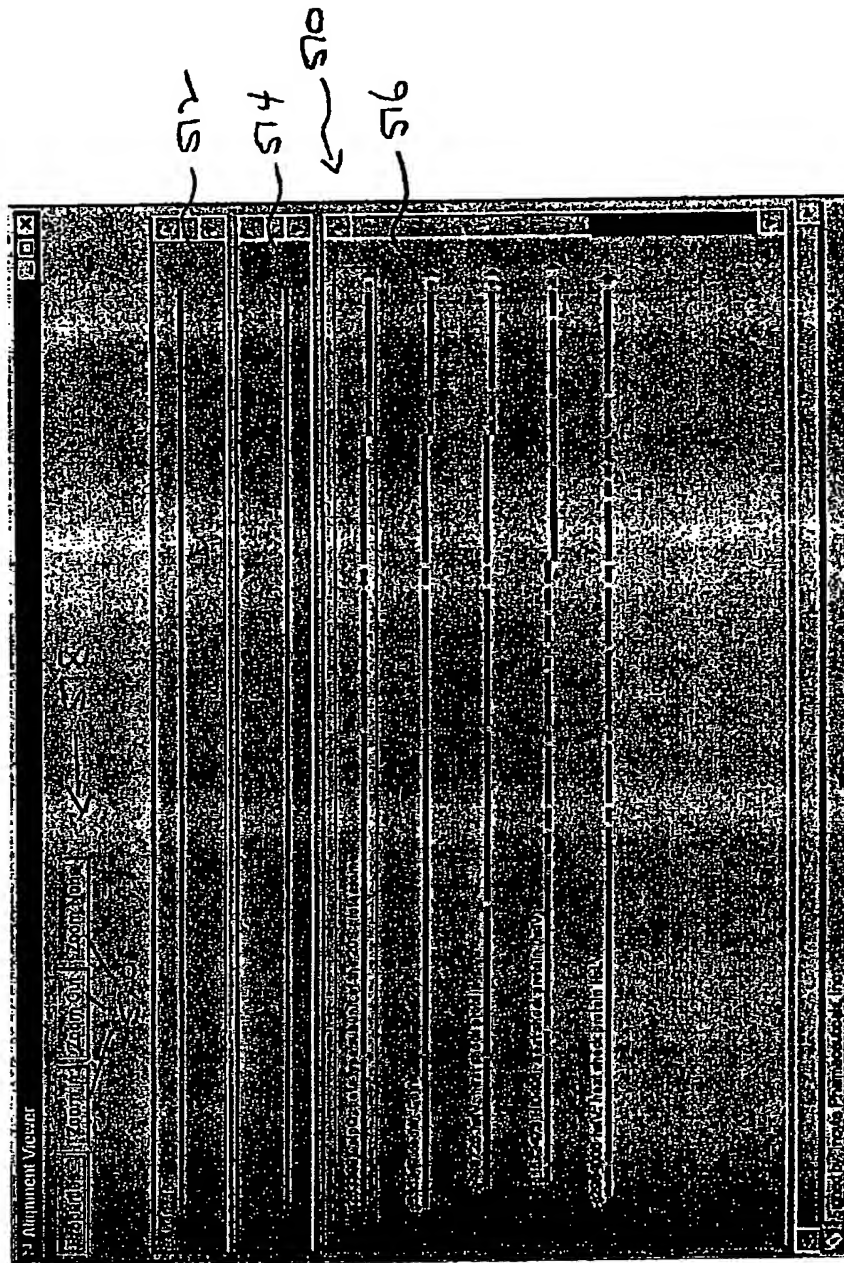


Fig. 5B

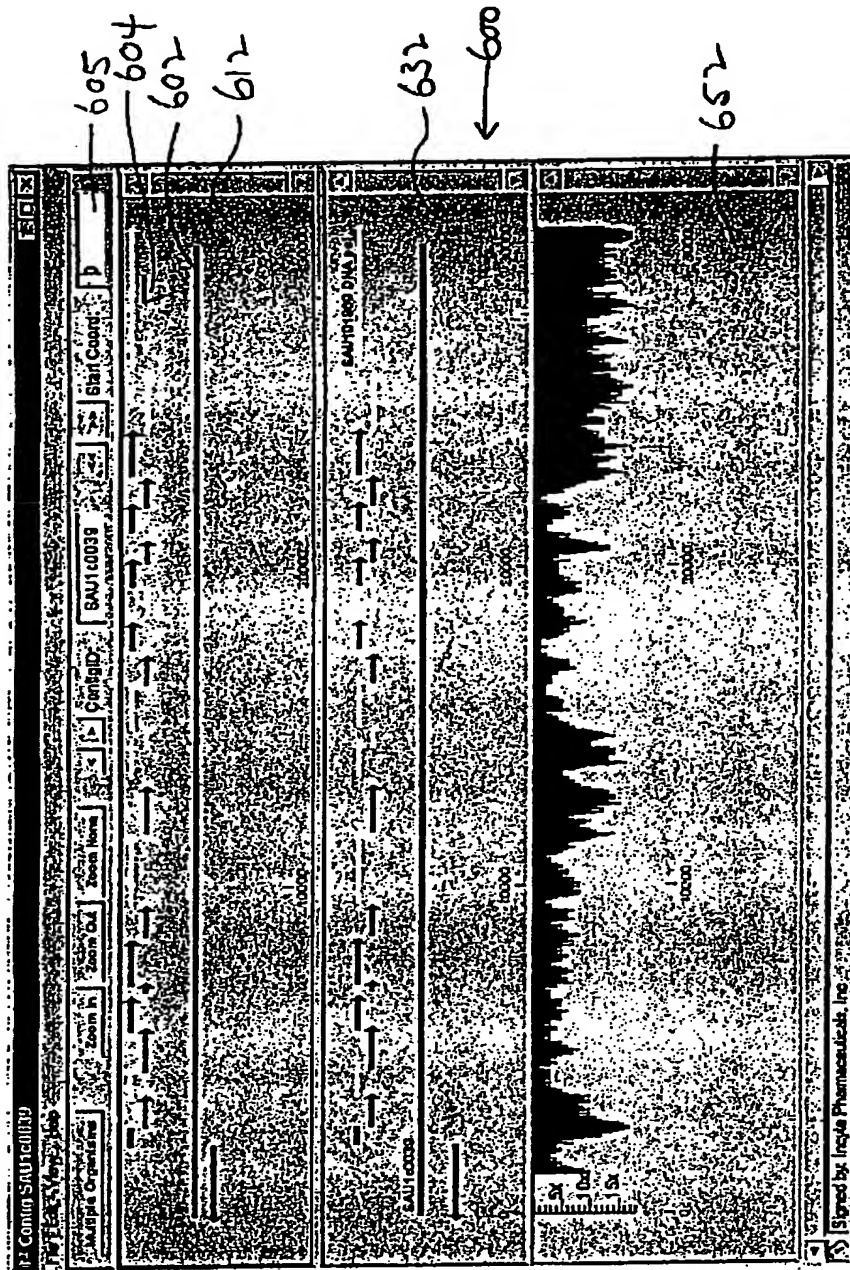


FIG. 6

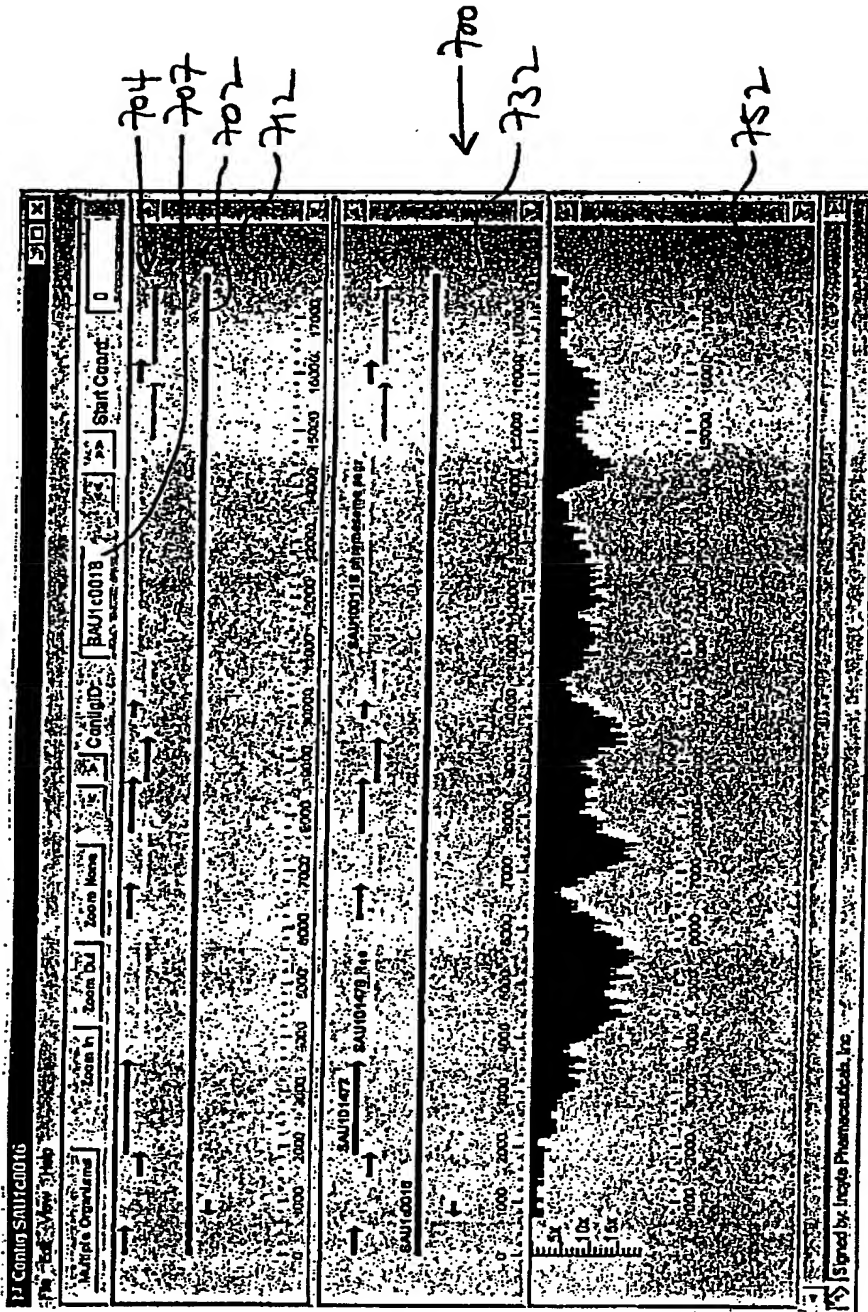


FIG. 7

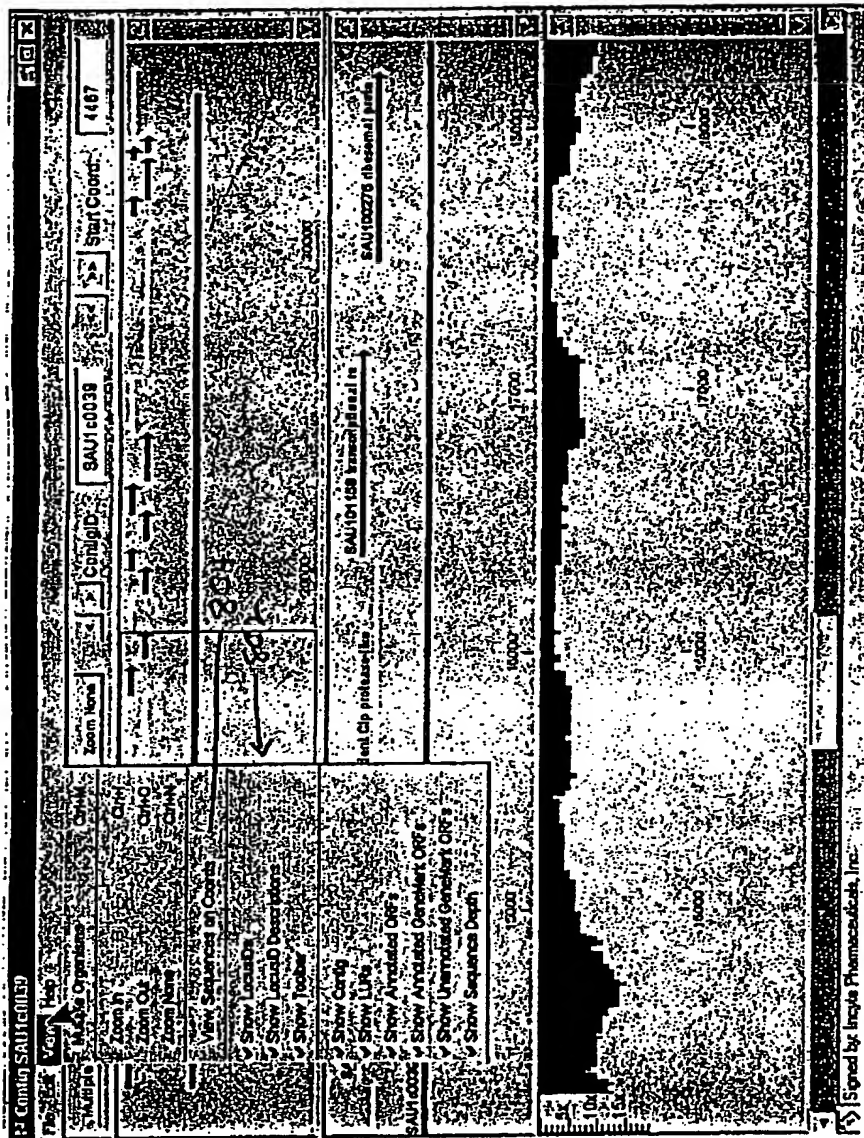


FIG. 8A

← 810

Sequence	Coordinates
808603456R1	(5142,5336)
808603054F1	(5201,5880)
808601824R1	(4288,5867)
808607074F1	(5607,6108)
808608114R1	(5888,6159)
801004114F1	(5848,8454)
801003848F1	(5848,8476)
808604249R1	(5901,8349)
808610122F1	(6080,8515)
808607288R1	(6238,8789)
801009208F1	(6388,8846)
808607038R1	(6615,7117)
808601448R1	(6780,7283)
808608588R1	(6803,7383)
808607551R1	(6882,7414)
808607074R1	(6897,7419)

← 812

814

FIG. 8B

56.80

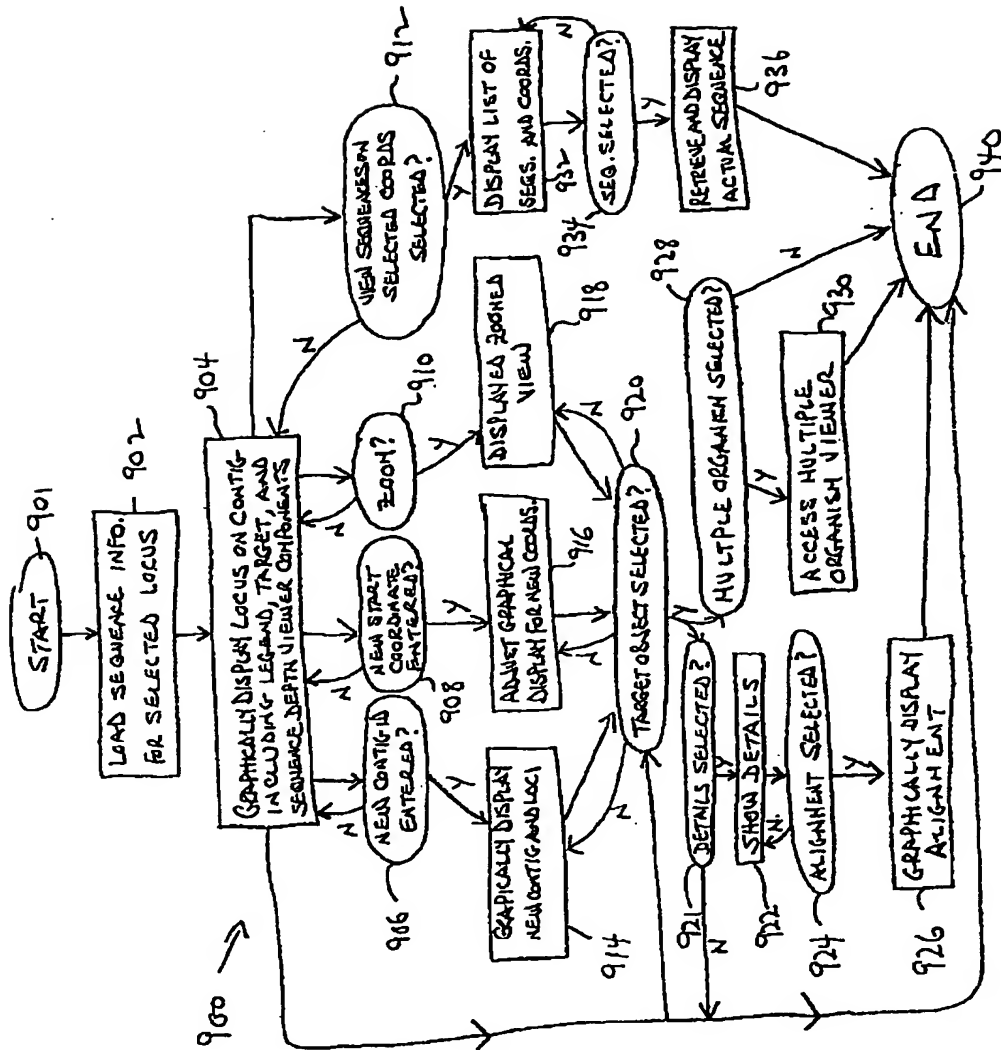


FIG. 9

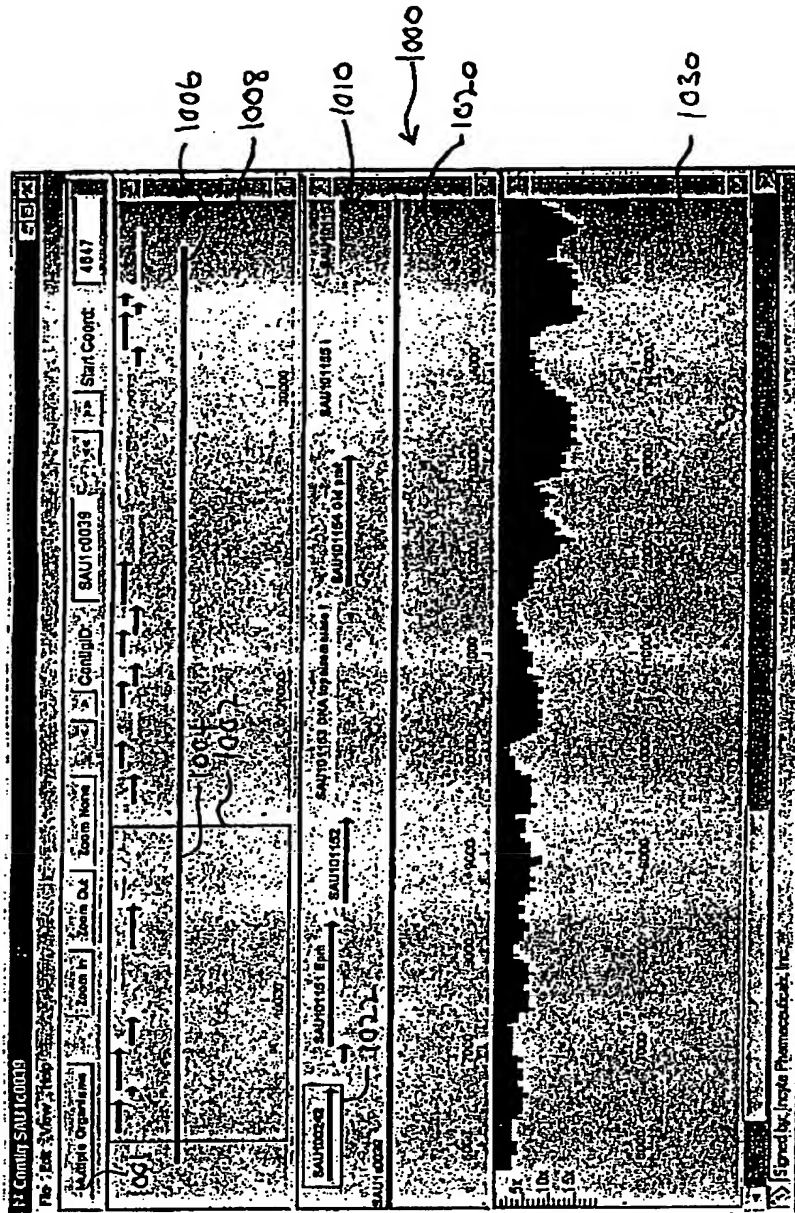


FIG. 10A

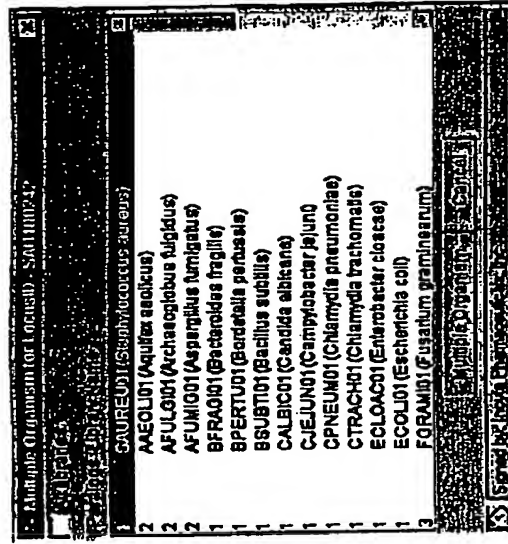


FIG. 108

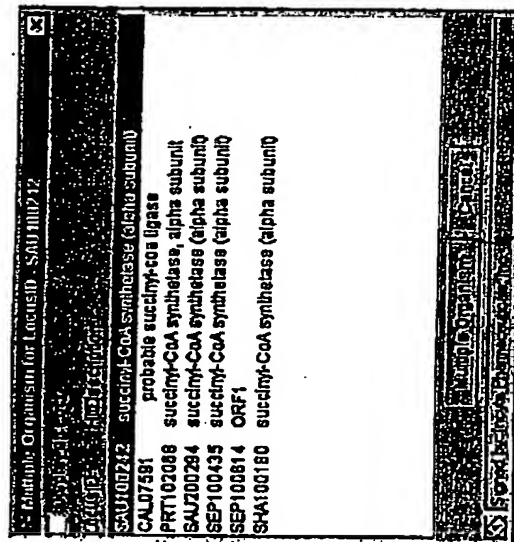


FIG. 10C

1040

1045

1042

1046

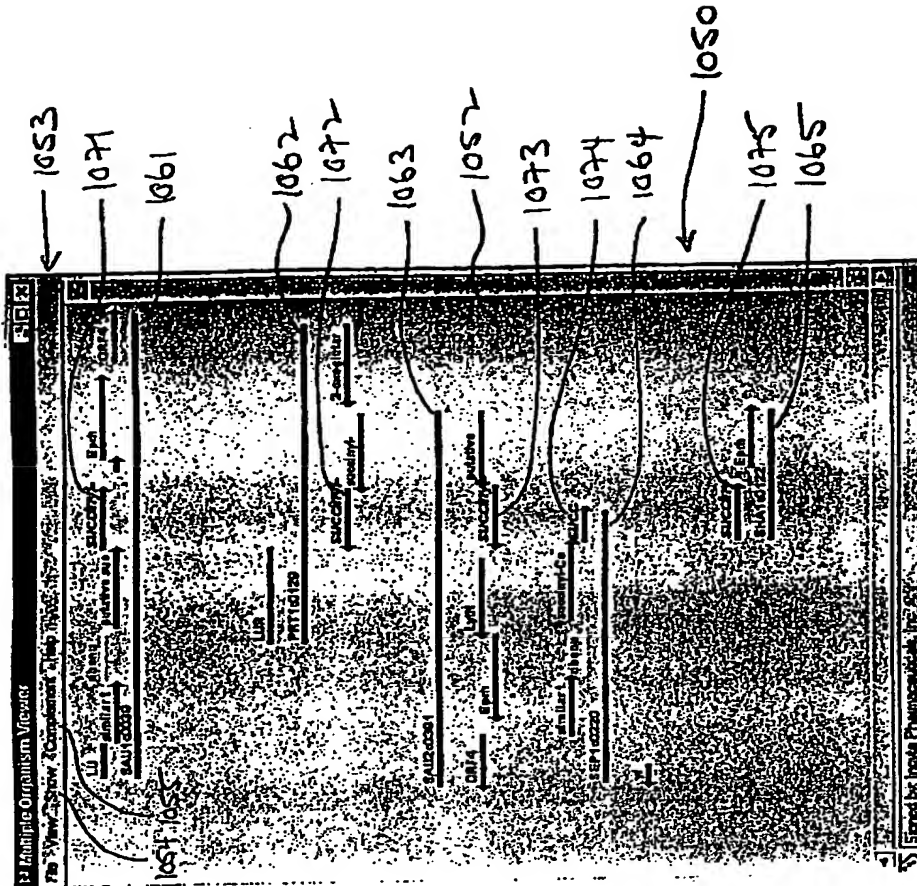


Fig. 10A

← 1080

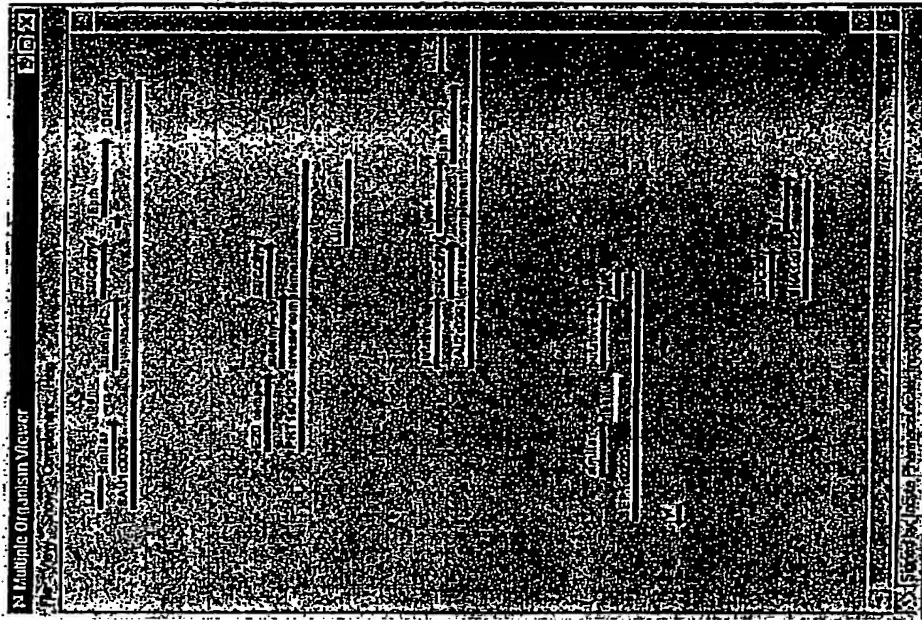


Fig. 10E

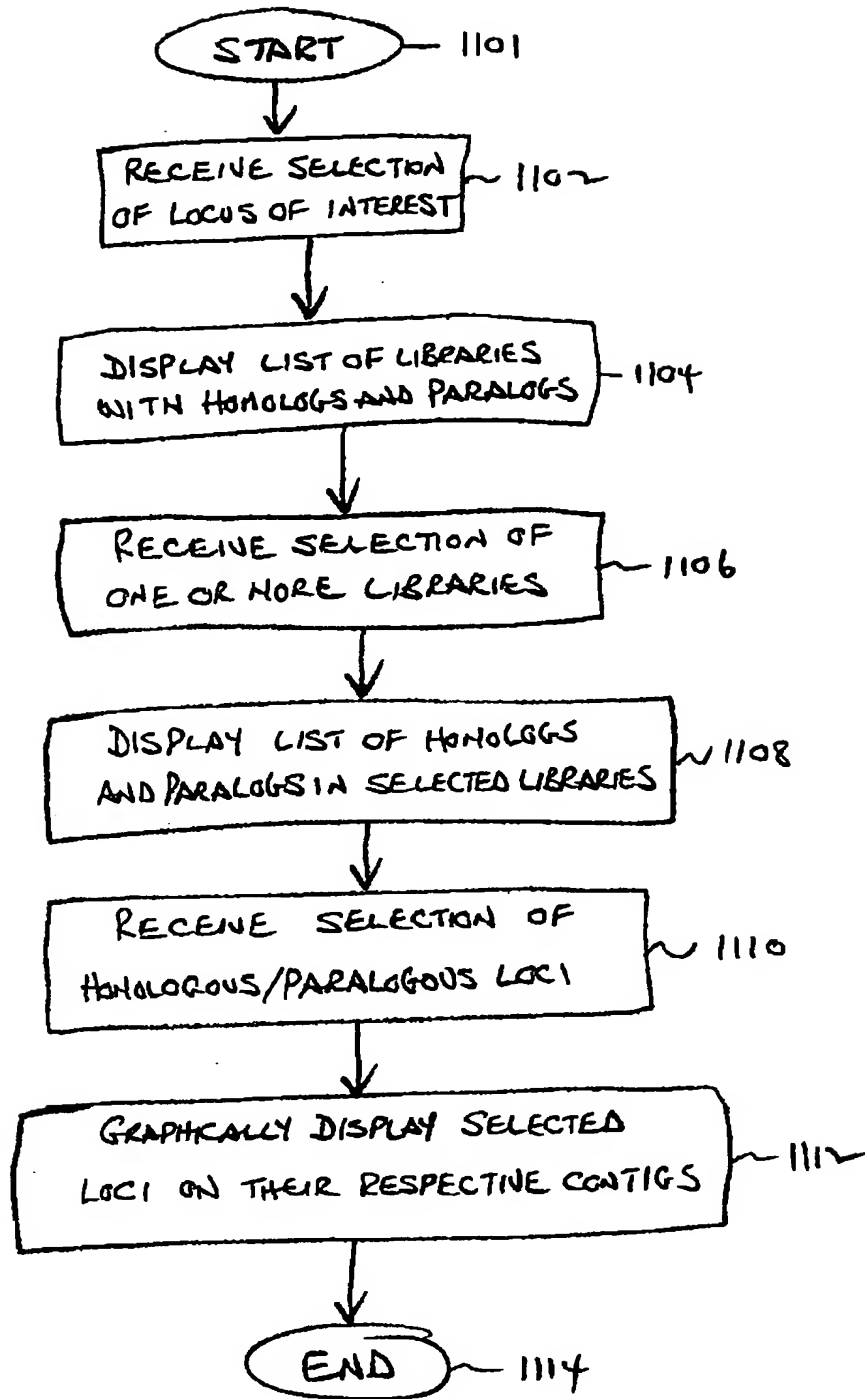


FIG. 11

1. Abstract

Disclosed are methods, media and systems for graphically displaying computer-based biomolecular sequence information. Generally, biomolecular sequence information may be graphically depicted in a variety of different forms in accordance with the present invention. The sequence information may be composed of nucleotide or amino acid sequence information or both. The graphical depictions may be in several different formats providing different information relating to the sequences, and may be displayed in one or more screens of a computer user interface.

2. Representative Drawing

Figure 4A

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER: _____**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.